

Sketches for Matrix Norms: Faster, Smaller and More General

Vladimir Braverman*
Johns Hopkins University

Stephen R. Chestnut†
ETH Zurich

Robert Krauthgamer‡
Weizmann Institute of Science

Lin F. Yang§
Johns Hopkins University

October 5, 2016

Abstract

We design new sketching algorithms for unitarily invariant matrix norms, including the Schatten p -norms $\|\cdot\|_{S_p}$, and obtain, as a by-product, streaming algorithms that approximate the norm of a matrix A presented as a turnstile data stream. The primary advantage of our streaming algorithms is that they are simpler and faster than previous algorithms, while requiring the same or less storage. Our three main results are a faster sketch for estimating $\|A\|_{S_p}$, a smaller-space $O(1)$ -pass sketch for $\|A\|_{S_p}$, and more general sketching technique that yields sublinear-space approximations for a wide class of matrix norms. These improvements are powered by dimensionality reduction techniques that are modern incarnations of the Johnson-Lindenstrauss Lemma [JL84]. When $p \geq 2$ is even or A is PSD, our fast one-pass algorithm approximates $\|A\|_{S_p}$ in optimal, $n^{2-4/p}$, space with $O(1)$ update time and $o(n^{2.4(1-2/p)})$ time to extract the approximation from the sketch, while the $\lceil p/2 \rceil$ -pass algorithm is built on a smaller sketch of size $n^{1-1/(p-1)}$ with $O(1)$ update time and $n^{1-1/(p-1)}$ query time. Finally, for a PSD matrix A and a unitarily invariant norm $l(\cdot)$, we prove that one can obtain an approximation to $l(A)$ from a sketch GAH^T where G and H are independent Oblivious Subspace Embeddings and the dimension of the sketch is polynomial in the intrinsic dimension of A . The intrinsic dimension of a matrix is a robust version of the rank that is equal to the ratio $\sum_i \sigma_i / \sigma_1$. It is small, e.g., for models in machine learning which consist of a low rank matrix plus noise. Naturally, this leads to much smaller sketches for many norms.

*Email: vova@cs.jhu.edu. This material is based upon work supported in part by the National Science Foundation under Grant No. 1447639, by the Google Faculty Award and by DARPA grant N660001-1-2-4014. Its contents are solely the responsibility of the authors and do not represent the official view of DARPA or the Department of Defense.

†Email: stephenc@ethz.ch

‡Email: robert.krauthgamer@weizmann.ac.il. Work supported in part by the Israel Science Foundation grant #897/13.

§Email: lyang@pha.jhu.edu

1 Introduction

In many applications, data is represented as a matrix $A \in \mathbb{R}^{m \times n}$ and a basic step in processing such data is to compute a norm $l(A)$. We focus on norms $l(A)$ that are unitarily invariant, which means they depend only on the singular values of the matrix.¹ Indeed, a fundamental characterization by John von Neumann from 1937 says that these are exactly the symmetric norms of the singular values.² Thus, for any unitarily invariant norm l on $\mathbb{R}^{m \times n}$ there is a symmetric norm l' on $\mathbb{R}^{\min(m,n)}$, such that $l(A) \equiv l'(\sigma_1, \dots, \sigma_{\min(m,n)})$, where $\sigma_1 \geq \dots \geq \sigma_{\min(m,n)}$ are the singular values of A . The class of unitarily invariant norms includes many important norms, like the spectral norm $\|A\|_2 = \sigma_1(A)$ and Frobenius norm $\|A\|_F = (\sum_{j \geq 1} \sigma_j^2)^{1/2}$, and more generally all Schatten p -norms $\|A\|_{S_p} = (\sum_{j \geq 1} \sigma_j^p)^{1/p}$ and Ky Fan k -norms $\|A\|_{(k)} = \sum_{j=1}^k \sigma_j$.

We study the *sketching model*, in which the input is mapped to a (possibly randomized) sketch, which suffices for computing the output. In our case of computing matrix norms, the input matrix A is mapped to a sketch $\mathbf{sk}(A)$, which in turn is used to approximate some norm $l(A)$. A *bilinear sketch* is a (randomized) function $\mathbf{sk} : A \mapsto GAH^T$ for some (random) matrices $G \in \mathbb{R}^{s \times m}$ and $H \in \mathbb{R}^{t \times n}$, and we say that $\mathbf{sk} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{s \times t}$ has dimension $s \cdot t$. Notice that a bilinear sketch is *linear* in the sense that $\mathbf{sk}(A + B) = \mathbf{sk}(A) + \mathbf{sk}(B)$ and $\mathbf{sk}(cA) = c\mathbf{sk}(A)$ for all matrices A, B and $c \in \mathbb{R}$. The main challenge in designing a sketch is to minimize the dimension. Unfortunately, computing $l(A)$ exactly typically requires dimension $\Omega(n^2)$, so we only ask the sketch to achieve, for a given $\epsilon > 0$, a $(1 \pm \epsilon)$ -approximation with probability $2/3$.

Streaming Matrices. Linear sketching is a major building block for streaming algorithms, because if the algorithm maintains a sketch of the stream seen so far, then the update procedure is straightforward from the linearity of the sketch. In our setting, the input matrix A is given as a stream of updates. At the start, the matrix A is initialized to zero, and each stream item of the form (i, j, δ) represents an update $A_{ij} \leftarrow A_{ij} + \delta$. This model captures many access patterns, like scanning the entries of A in row-order, column-order, or any fixed or dynamic order; or a sparse matrix presented as a list of its non-zero entries, written as tuples (i, j, A_{ij}) in any order. We consider the *turnstile* model, where $\delta \in \mathbb{R}$ is not restricted.

Updating a bilinear sketch $\mathbf{sk} : A \mapsto GAH^T$ is straightforward. Indeed, if e_i denotes the i -th standard basis column vector, then $\mathbf{sk}(A + \delta e_i e_j^T) = \mathbf{sk}(A) + \delta \mathbf{sk}(e_i e_j^T) = \mathbf{sk}(A) + \delta(Ge_i)(He_j)^T$, hence an update amounts to adding to the current sketch a rank-one matrix. This would usually take $O(st)$ time, but in some cases it can be faster — if every column of G and H has at most B non-zeros, then the increment matrix has only B^2 non-zeros, and such an update can usually be implemented in $O(B^2)$ time. When the stream ends, which is called the query phase, the algorithm computes its output from the sketch (only).

The sketching model is very attractive for analyzing large matrices, which are commonplace in current data sets. Storing $\mathbf{sk}(A)$ obviously requires less space than A , which can be crucial to speeding up data transfer between machines or even inside a machine. Perhaps more importantly, sketching can directly improve runtime, because matrix operations, like multiplication or computing singular values, are faster on a smaller matrix $\mathbf{sk}(A)$ than on A . Indeed, faster algorithms for many

¹Formally, a matrix norm $l : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}_+$ is *unitarily invariant* if $l(UAV) = l(A)$ for all orthogonal matrices U, V (of order m and n , respectively).

²A norm on \mathbb{R}^N is called symmetric (or a symmetric gauge function) if it is invariant under sign-flips and coordinate-permutations, see e.g. [Bha97, Chapter IV].

Problem: $\ A\ _{S_p}$ for PSD A , integer $p \geq 2$ (or general A , even p)				
passes	space	update time	query time	
1	$\epsilon^{-2}n^{2-4/p}$	$\epsilon^{-2}n^{2-4/p}$	$\epsilon^{-2}n^{p-2}$	[LNW14]
1	$\epsilon^{-2}n^{2-4/p}$	ϵ^{-2}	$\epsilon^{-2}n^{(1-2/p)\omega}$	Theorems 3.2 and A.11
$\lceil p/2 \rceil$	$\epsilon^{-2}n$	ϵ^{-2}	$\epsilon^{-2}n$	[Woo14, Theorem 6.1]
$\lceil p/2 \rceil$	$\epsilon^{-2}n^{1-1/(p-1)}$	ϵ^{-2}	$\epsilon^{-2}n^{(1-1/(p-1))}$	Theorems 3.5 and A.11

Problem: $l(A)$ for PSD A when $\text{intdim}(A) \leq r$ and $l(I_d) \leq d^{1-\alpha}l(I_1)$				
passes	space	update time	query time	
1	$(\epsilon^{-2}r)^{2\alpha} \text{polylog } n$	$\epsilon^{-2} \log^2 n$	depends on l	Proposition 4.2

Table 1: Streaming algorithms for $(1 + \epsilon)$ -approximation of various matrix norms. These bounds count words of space, and omit $O_p(1)$ factors (i.e., constants assuming fixed p).

Numerical Linear Algebra tasks, from regression to spectral sparsification, were recently designed using sketching (of matrices and of vectors), see Woodruff’s survey [Woo14] and references therein. In many of these algorithms, the overall runtime is proportional to the number of non-zeros in A (instead of the dimensions of A), which becomes the bottleneck and calls for a sketch with a fast update time.

Our Contribution in a Nutshell. We design new sketching algorithms for unitarily invariant matrix norms, and obtain as a (sometimes less immediate) by-product streaming algorithms that compute these norms in the turnstile model. The primary advantage of our streaming algorithms is that they are simpler and faster than previous algorithms, while requiring the same or less storage. These improvements are powered by dimension reduction techniques that are modern incarnations of the Johnson-Lindenstrauss Lemma [JL84].

Our main and most technical result is for computing Schatten p -norms. But along the way we design algorithms with several new features compared to previous work. First, we show that the storage requirement can be related to the input’s *intrinsic dimension*, defined as $(\sum_{j \geq 1} \sigma_j)/\sigma_1$. Second, our results cover many unitarily invariant matrix norms, while previous algorithms and lower bounds focused on top singular values and/or Schatten norms. Third, our algorithms rely on now-standard primitives, and are thus be faster (as mentioned above) and also easier to implement and to build upon.

1.1 Schatten p -norms

Our main result is a new method for estimating $\|A\|_{S_p}$, the Schatten p -norm of a PSD matrix $A \in \mathbb{R}^{n \times n}$ for integer $p \geq 2$. This method yields two new streaming algorithms, which require, respectively, one pass and $\lceil p/2 \rceil$ passes over the input. Both algorithms are at least as good as the previous ones in all three standard performance measures of storage, update time, and query time; and each algorithm offers significant improvements in two out of these three. We provide a detailed comparison in Table 1, omitting for simplicity factors that depend on p and ϵ , and briefly discuss now only the highlights. Our one-pass algorithm achieves update time $O(1)$ compared with the previous $\text{poly}(n)$, and query time $O(n^{\omega(1-p/2)})$, where $\omega \leq 2.373$ is the matrix multiplication

exponent [Le 14], compared with the previous n^{p-2} . And our multi-pass algorithm requires storage that is sublinear in n , compared with $O(n)$ previously. We note that if p is even, then the above results extend to arbitrary $A \in \mathbb{R}^{m \times n}$ by a standard argument.

Our technical innovation is an unbiased estimator of $\text{Tr}(A^p)$ for a *symmetric* (and not only PSD) matrix $A \in \mathbb{R}^{n \times n}$. To see why this is useful, denote the eigenvalues of A by $\lambda_1 \geq \dots \geq \lambda_n$, and observe that if A is PSD (or alternatively if p is even), then $\text{Tr}(A^p) = \sum_i \lambda_i^p = \sum_i \sigma_i(A)^p = \|A\|_p^p$. Our estimator has the form

$$X := \text{Tr}(G_1 A G_2^T G_2 A G_3^T \dots G_p A G_1^T), \quad (1)$$

where $G_i \in \mathbb{R}^{t \times n}$ are certain random matrices.³ This estimator X can be computed from the p bilinear sketches $\{G_i A G_{i+1}^T\}_{i \in [p]}$, where by convention, $i_{p+1} := 1$, by straightforward matrix multiplication. And if say $t = O(n^{1-2/p})$, then each bilinear sketch has dimension $O(t^2) = O(n^{2-4/p})$. These determine the streaming algorithm's storage requirement and query time, and clearly, if the matrices $\{G_i\}_{i \in [p]}$ have sparse columns, then the updates will be fast.

The main difficulty is to bound the estimator's variance, which highly depends on the choice of the matrices $\{G_i\}_{i \in [p]}$. The basics of this technique can be seen in the case $p = 4$, if the G_i 's satisfy the following definition.

Definition 1.1. An (ϵ, δ, d) -Johnson-Lindenstrauss Transformation (JLT) is a random matrix $S \in \mathbb{R}^{t \times n}$, such that for every $V \subseteq \mathbb{R}^n$ of cardinality $|V| \leq d$,

$$\Pr \left[\forall x \in V, \|Sx\|_2^2 \in (1 \pm \epsilon) \|x\|_2^2 \right] \geq 1 - \delta.$$

An (ϵ, δ, d) -JLT can be constructed with $t = O(\epsilon^{-2} \log(d/\delta))$ rows, which is optimal [JW13]. While using independent Gaussian entries $N(0, 1/t)$ works, there is a construction with only $O(\epsilon^{-1} \log(1/\delta))$ non-zero entries per column [KN14].

The case $p = 4$ has a particularly short and simple analysis, whenever G_1 and G_2 are independent (ϵ, δ, n) -JLT matrices, which we can achieve with $t = O(\epsilon^{-2} \log n)$. The first idea is to “peel off” G_i from both sides, using that for any PSD matrix M , with high probability $\text{Tr}(G_i M G_i^T) \in (1 \pm \epsilon) \text{Tr}(M)$ (see Lemma A.1 for a precise statement). A second idea is to use the identity $\text{Tr}(XY) = \sum_{ij} X_{ij} Y_{ij} = \text{Tr}(YX)$ to rewrite $\text{Tr}(A A^T G_2^T G_2 A A^T) = \text{Tr}(G_2 A A^T A A^T G_2^T)$. Now using the first idea once again, we are likely to arrive at an approximation to $\text{Tr}(A A^T A A^T) = \|A\|_4^4$. The full details are given in Section 3.1.

The sketching method extends from $p = 4$ to any integer $p \geq 2$, but the simple analysis above breaks (because for $p > 4$ the “inside” matrix M is no longer PSD) and thus our analysis is much more involved. We first analyze G_i 's with independent Gaussian entries, by a careful expansion of the fourth moment of X , which exploits certain cancellations occurring (only) for Gaussians. We then consider G_i 's that are sampled from a particular sparse JLT due to [TZ04], and employ a symmetrization-and-decoupling argument to compare the variance of X in this case with that of Gaussian G_i 's.

We make two technical remarks. First, proving $\mathbb{E}[X] = \text{Tr}(A^p)$ is straightforward. Indeed, by the second idea above, we can rewrite our $X = \text{Tr}(G_1 A G_2^T G_2 A G_3^T \dots G_p A G_1^T)$ as $\text{Tr}(G_1^T G_1 A G_2^T G_2 A \dots G_p^T G_p A)$. Now using $\mathbb{E}[G_i^T G_i] = I$ together with linearity of the trace and of expectation, we get that

³After posting our result on the arXiv, we became aware Yi Li [Li11] had analyzed a similar estimator for Schatten p -norm, where the G_i are random sign matrices.

$\mathbb{E}[X] = \text{Tr}(A^p)$. Second, after setting $t = O(n^{1-2/p})$ (independent of ϵ), our bound on the variance is $O(\mathbb{E}[X])^2$, which we can decrease by standard $O(1/\epsilon^2)$ repetitions. See Sections 3.2 and 3.4 for details.

The multi-pass streaming algorithm is implemented slightly differently, in that $G_1 \in \mathbb{R}^{1 \times n}$, i.e., has only one row. The other matrices $G_2, \dots, G_p \in \mathbb{R}^{t \times n}$ are as before, although we now set $t = O(n^{1-1/(p-1)})$. Our estimator X can be computed in $\lceil p/2 \rceil$ passes with space only $2t$ as follows. On the first pass, compute vectors $X_L \leftarrow G_1 A G_2^T \in \mathbb{R}^{1 \times t}$ and $X_R \leftarrow G_p^T A G_1 \in \mathbb{R}^{t \times 1}$, and then on the i -th pass update $X_L \leftarrow X_L G_i^T A G_{i+1}$ and $X_R \leftarrow G_{p-i+1} A G_{p-i+2}^T X_R$. Notice that the computation in each pass is linear in A . For even p , after completing $p/2$ passes, compute and output $X' = X_L X_R \in \mathbb{R}$ (and similarly for odd p). This X' is similar to the estimator X from above, except for the new dimensions of the G_i 's. See Sections 3.3 and 3.4 for details.

This multi-pass algorithm offers a very significant space saving over the one-pass algorithm. It is also a bit surprising because it is getting close to the corresponding vector norm, namely, ℓ_p -norm on \mathbb{R}^n , for which the optimal space for $O(p)$ passes is $\tilde{O}(n^{1-2/p})$ bits. In fact, for the vector norm, $O(p)$ passes do not significantly reduce the storage needed compared with one pass, which stands in sharp contrast to Schatten p -norm. As before, if p is even then the algorithm extends to arbitrary $A \in \mathbb{R}^{m \times n}$ by a standard argument.

1.2 Unitarily Invariant Norms

The second dimension of this paper is a sketching algorithm for unitarily invariant matrix norms. This sketch is limited to PSD input matrices, but it is even simpler than the sketch for Schatten p -norms — one simply computes a single bilinear form GAH^T for two independent matrices $G, H \in \mathbb{R}^{t \times n}$ that satisfy the following definition.

Definition 1.2. An (ϵ, δ, d) -Oblivious Subspace Embedding (OSE) is a random matrix $S \in \mathbb{R}^{t \times n}$, such that for every matrix $A \in \mathbb{R}^{n \times d}$,

$$\Pr \left[\forall x \in \mathbb{R}^d, \|SAx\|_2 \in (1 \pm \epsilon) \|Ax\|_2 \right] \geq 1 - \delta.$$

Given a unitarily invariant norm $l : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, we approximate $l(A)$ from the sketch simply by $l(GAH^T)$. The sketching dimension is t^2 , where t depends on the structure of l and the *intrinsic dimension* of A , defined as $\text{intdim}(A) := (\sum_{j \geq 1} \sigma_j) / \sigma_1$ [Tro15]. At a high level, we use $t = (\text{intdim}(A)/\epsilon)^{O(1)}$, see Section 4 for details. The motivation to understand the problem in terms of $\text{intdim}(A)$ is to get around the strong lower bounds that are known for sketching many matrix norms. For example, every one-pass approximation of the spectral norm of a general matrix requires $\Omega(n^2)$ bits [LNW14], but by our Theorem 4.1, a sketch of dimension $O(\epsilon^{-2} \text{intdim}(A)^2)$ suffices. The intrinsic dimension is at most the rank of the matrix, so this is a weaker assumption than small rank. Intrinsic dimension can be much smaller than rank, which leads to smaller sketches. This improvement can be important for applications. For instance, a common model assumes the data are column vectors from a low dimensional subspace of \mathbb{R}^n that have been corrupted by a small amount of noise. The data matrix, then, will have small intrinsic dimension but full rank.

Using similar techniques, we show that also the top- k singular values of A can be recovered from bilinear sketches of the form GAG^T or GAH^T , up to multiplicative error $1 \pm \epsilon$ and additive error $\sum_{i=k+1}^n |\sigma_i|/k$ or $(\sum_{i=k+1}^n \sigma_i^2/k)^{1/2}$. These are the same error bounds as in [AN13], except that our G and H are OSE matrices instead of Gaussians. We can thus improve the update time

to $O(\log^2 n/\epsilon^2)$, at the expense of slightly increasing the sketch size (by logarithmic factors). The details appear in Section 5.

Sketching based on OSEs provide flexibility to the user, who has a choice among several OSE distributions, with different useful properties like sparsity [NN13, KN14], fast multiplication routines (e.g., based on the Fast Fourier Transform) [AC09, AL13, CW13], and compressed storage of the OSE matrix and limited randomness [KMN11].

1.3 Previous work

The aforementioned algorithm of [LNW14] uses a single sketching matrix G , for example, if A is PSD, then their sketch is $S = GAG^T$, where $G \in \mathbb{R}^{t \times n}$ is a Gaussian matrix. Its estimate for $\|A\|_{S_p}$ is produced by summing over all “cycles” $S_{i_1, i_2} S_{i_2, i_3} \dots S_{i_p, i_1}$, where $i_1, \dots, i_p \in [t]$ are distinct. Our sketch improves over theirs in both update time and query time. The only other streaming algorithm for Schatten p -norm that we are aware of is that of [LW16a, Theorem 7], uses space $O(n^{1-2/p} \text{poly}(1/\epsilon, \log n))$ but works only for matrices that have $O(1)$ -entries per row and per column.

Several strong lower bounds are known for approximating Schatten p -norm and other matrix functions, both for the dimension of a sketch and for storage requirement (bits). Li, Nguyen and Woodruff [LNW14] prove that every bilinear sketch that can approximate rank and Schatten p -norm for $0 \leq p < 2$ must have dimension $\Omega(\sqrt{n})$. They also show an $\Omega(n^{1-\epsilon})$ lower bound for Schatten 1-norm. Li and Woodruff [LW16b] show that every linear sketch for Schatten p -norm, $p \geq 2$, requires dimension $\Omega(n^{2-4/p})$. In [LW16a], they prove space complexity lower bounds that hold even when the input matrix has $O(1)$ -entries per row and per column. Specifically, they show that one-pass streaming algorithms that $(1 \pm \epsilon)$ -approximates various functions of the singular values, including Schatten p -norms when p is not an even integer, require $\Omega(n^{1-f(\epsilon)})$ bits of space for some function $f(x) \rightarrow 0$ as $x \rightarrow 0$. Additional space lower bounds, e.g., for $p \in [1, 2)$, can be deduced from a general statement of [AKR15], see [LW16a, Table 1] for details.

2 Notation and Preliminaries

Our space bounds are stated in terms of sketch dimension (number of entries). The number of bits required can be larger by a $\log nM$ factor, where M is the absolute ratio of the largest element in the matrix to the smallest. We call a matrix a *Gaussian matrix* if its entries are independent $N(0, 1)$ random variables. A matrix G of dimension $t \times n$ is a *column-normalized* Gaussian matrix if $G = G'/\sqrt{t}$, where G' is a Gaussian matrix. Now-standard techniques such as Nisan’s Pseudo-random generator or k -wise independence can be used to derandomize Gaussian matrices for use in sketching algorithms. Column-normalized Gaussian matrices serve as JLTs and OSEs. In particular, there exists a constant c such that if G be a $t \times n$ column-normalized Gaussian matrix with $t \geq \frac{c}{\epsilon^2} \log \frac{d}{\delta}$, then G is a (ϵ, δ, d) -JLT [IM98], and if $t \geq \frac{c}{\epsilon^2} (d + \log \frac{1}{\delta})$, then G is a (ϵ, δ, d) -OSE [Woo14]. Recently, Cohen showed a nearly-tight upper bound for a sparse OSE.

Theorem 2.1 ([Coh16]). *For every $B > 1$ there exists an (ϵ, δ, d) -OSE matrix of dimension $m \times n$, where $m = O(\epsilon^{-2} B d \log(d/\delta))$ and each column of the OSE matrix has at most $O(\epsilon^{-1} \log_B(d/\delta))$ non-zero entries.*

The next lemma shows that the OSE property is preserved upon right multiplication by a matrix

with orthonormal columns. We later use this lemma to diagonalize a symmetric matrix A that is within the bilinear sketch, effectively reducing to the simpler case where A is a diagonal matrix.

Lemma 2.2. *Let $S \in \mathbb{R}^{t \times n}$ be an (ϵ, δ, d) -OSE matrix, and let $U \in \mathbb{R}^{n \times r}$ be a matrix with orthonormal columns. Then SU is an $(\epsilon, \delta, \min(r, d))$ -OSE matrix (for the space \mathbb{R}^r).*

Proof. Consider $A \in \mathbb{R}^{r \times \min\{r, d\}}$. By applying the OSE guarantee of S to UA (since the OSE guarantee extends to every dimension $d' \leq d$), we get that with probability at least $1 - \delta$,

$$\forall x \in \mathbb{R}^{\min\{r, d\}}, \quad \|S(UA)x\|_2 \in (1 \pm \epsilon)\|(UA)x\|_2 = (1 \pm \epsilon)\|Ax\|_2,$$

where the last equality is because of the orthonormal columns of U , which imply that for all $y \in \mathbb{R}^r$, we have $\|Uy\|_2 = \|y\|_2$. \square

3 Schatten p -norm

The main result in this section is a new one-pass streaming algorithm for estimating the Schatten p -norm, for integer $p \geq 2$. When p is odd, it additionally requires that the input matrix is PSD. The first version of this algorithm, described in Section 3.2, has the same storage requirement of $\tilde{O}_p(n^{2-4/p}/\epsilon^2)$ bits as the previous algorithm of [LNW14] that uses cycle sums, but has simpler analysis and faster query time, which is roughly matrix multiplication time n^ω instead of n^p . Moreover, it is based on a new method that leads to a $\lceil p/2 \rceil$ -pass algorithm with storage requirement $\tilde{O}_p(n^{1-1/(p-1)}/\epsilon^2)$ bits, as described in Section 3.3. Previously, the algorithm in [Woo14, Theorem 6.1] has the same number of passes but larger storage requirement $O(n/\epsilon^2)$.⁴ Finally, we improve the update time, as described in Section 3.4, by employing the sketching matrices G_i that are certain sparse matrices instead of Gaussians.

We start in Section 3.1 with the case $p = 4$, which is based on the same sketch but is significantly easier to analyze.

3.1 Schatten-4 Norm using JLT matrices

Theorem 3.1. *Let $G_1, G_2 \in \mathbb{R}^{t \times n}$ be independent $(\epsilon, \delta/n, 1)$ -JLT matrices. Then for every $A \in \mathbb{R}^{n \times m}$,*

$$\Pr \left[\text{Tr}(G_1 A A^T G_2^T G_2 A A^T G_1^T) \in (1 \pm 2\epsilon)^2 \|A\|_{S_4}^4 \right] = 1 - 2\delta.$$

Thus, one can find a $(1 \pm \epsilon)$ -approximation to the Schatten-4 norm of a general matrix $A \in \mathbb{R}^{n \times m}$ using a linear sketch of dimension $O(\epsilon^{-2} n \log n)$.

The proof of this Theorem appears in Section A.1. If each column of G_i has only s non-zero entries, it is easy to see that the update time of this linear sketch is $O(s)$, assuming any entry of G_1 and G_2 can be accessed in $O(1)$ time (in a streaming algorithm, the entries are usually computed from a small random seed in $\text{polylog}(n)$ time). The query time is dominated by multiplying a matrix of size $t \times n$ with one of size $n \times t$, and thus take $O(t^\omega \cdot n/t) = \tilde{O}(n^\omega / \epsilon^{2(\omega-1)})$.

⁴We note that also in [Woo14, Theorem 6.1] it is required that p is even or that the input matrix is PSD, but this is erroneously omitted.

3.2 Schatten p -norm Using Gaussians

We now design a sketch for Schatten- p norm that uses column-normalized Gaussian matrices. We will later extend and refine it to improve the per-update processing time.

Theorem 3.2. *For every $0 < \epsilon < 1/2$ and integer $p \geq 2$, there is an algorithm that outputs at $(1 \pm \epsilon)$ -approximation to the Schatten- p norm of a PSD matrix $A \in \mathbb{R}^{n \times n}$ using a randomized linear sketch of dimension $s = O_p(\epsilon^{-2} n^{2-4/p})$. The update time (for each entry in A) is $O(s)$ and the query time (for computing the estimate) is $O(\epsilon^{-2} n^{(1-2/p)\omega})$, where $\omega < 2.373$ is the matrix multiplication constant.*

If p is even, the above algorithm extends to a general matrix $A \in \mathbb{R}^{n \times m}$.

The first part of the theorem (for PSD matrices) follows directly from Proposition 3.3 below. The proposition is applicable to all symmetric matrices, but $\|A\|_{S_p}^p = \text{Tr}(A^p)$ only for PSD matrices or even p . The linear sketch stores $G_i A G_{i+1}^T$ for $i = 1, \dots, p$, where by convention $G_{p+1} = G_1$, repeated independently in parallel $O_p(1/\epsilon^2)$ times. Thus, the sketch has dimension $O_p(\epsilon^{-2} t^2)$. The estimator is obtained by computing the $O_p(1/\epsilon^2)$ independent copies of X and reporting their average. To analyze its accuracy, notice that a PSD matrix A satisfies $\mathbb{E}[X] = \text{Tr}(A^p) = \|A\|_p^p$. Then setting $t = n^{1-2/p}$ gives $\text{Var}(X) \leq O_p(\|A\|_{S_p}^{2p})$ and averaging multiple independent copies of X reduces the variance.

The second part (for general matrices), follows by using the same sketch, for the symmetric (but not PSD) matrix $B = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, because the nonzero singular values of B are those of A repeated twice, and $\|B\|_{S_p}^p = 2\|A\|_{S_p}^p = 2\text{Tr}(A^p)$, where the last equality uses that p is even.

Because the correctness of the algorithm comes by bounding the variance of X , it is enough that the entries in each Gaussian matrix are four-wise independent, which is crucial for applications with limited storage like streaming.

Proposition 3.3. *For integer $p \geq 2$ and $t \geq 1$, let $G_1, \dots, G_p \in \mathbb{R}^{t \times n}$ be independent column-normalized Gaussian matrices. Then for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, the estimator $X = \text{Tr}(G_1 A G_2^T G_2 A \dots G_p^T G_p A G_1^T)$ satisfies*

$$\mathbb{E}[X] = \text{Tr}(A^p) \text{ and } \text{Var}(X) = O_p \left(1 + \sum_{z=2}^{\lfloor p/2 \rfloor + 1} \left(\frac{n^{1-2/p}}{t} \right)^z + \sum_{z=2}^p \left(\frac{n^{1-2/z}}{t} \right)^z \right) \|A\|_{S_p}^{2p}.$$

The full proof of this proposition appears in Appendix A. We outline the general idea here. It is standard that Gaussian matrix is rotational invariant, i.e., G and GU are identically distributed for any orthogonal matrix U . Thus, by the Spectral Theorem, instead of considering symmetric matrix $A = U\Lambda U^T$, we can consider only its diagonalization Λ .

The proof of this proposition proceeds first by expanding X in terms of inner products of columns of the matrix G , i.e., $X = \sum_{i_1, i_2, \dots, i_p \in [n]} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_p} \langle g_{i_1}^{(1)}, g_{i_2}^{(1)} \rangle \langle g_{i_2}^{(2)}, g_{i_3}^{(2)} \rangle \dots \langle g_{i_p}^{(p)}, g_{i_1}^{(p)} \rangle$, where λ_i is the i -th eigenvalue of A and $g_{i_j}^{(j)}$ is the i_j -th column of G_j . We then expand $\mathbb{E}(X^2)$. The non-zero terms in $\mathbb{E}(X^2)$ are composed by only those terms of even powers in every eigenvalue. Computing the expectation of each term is straightforward because the entries of G are independent Gaussian random variables, but the crux of the proof is in bounding the sum of the terms. We introduce a collection of diagrams that aid in enumerating the terms according to their structure and computing the sum.

3.3 Multipass Algorithm

The proof of Proposition 3.3 relies on the matrices G_i being Gaussians in two places. First, we used it assume that the matrix A is diagonal, and in general we need to consider $G_i U$ instead of G_i . Second, the columns of these matrices have small variance/moments, as described in (6)-(7). We now generalize the proof to relax these requirements (e.g., to 4-wise independence) and obtain a multipass algorithm.

Lemma 3.4. *For integers $p \geq 2$ and $1 \leq t' \leq t$, let $G_1 \in \mathbb{R}^{t' \times n}$ and $G_2, \dots, G_p \in \mathbb{R}^{t \times n}$ be independent column-normalized Gaussian matrices with 4-wise independent entries. The for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, the estimator $X = \text{Tr}(G_1 A G_2^T G_2 A \dots G_p^T G_p A G_1^T)$ satisfies*

$$\mathbb{E}[X] = \text{Tr}(A^p) \text{ and } \text{Var}(X) = O_p \left(1 + \sum_{z=2}^{\lfloor p/2 \rfloor} \frac{n^{z-1-2(z-1)/p}}{t' t^{z-1}} + \sum_{z=2}^p \frac{n^{z-2}}{t' t^{z-1}} \right) \|A\|_{S_p}^{2p}.$$

The proof of this lemma appears in Appendix A. It is a direct corollary of the proof of Proposition 3.3, except that t' , the size of the first sketch matrix, is emphasized.

We can now use the above sketch to approximate the Schatten p -norm using $\tilde{O}(n^{1-1/(p-1)})$ bits of space with $\lceil p/2 \rceil$ passes over the input.

Theorem 3.5. *Let $p \geq 2$ be an even integer. There is a $\lceil p/2 \rceil$ -pass streaming algorithm, that on input matrix $A \in \mathbb{R}^{n \times m}$ with $n \geq m$ given as a stream, outputs an estimate X such that with probability at least 0.9,*

$$X \in (1 \pm \epsilon) \|A\|_{S_p}^p,$$

and uses $O_p(n^{1-1/(p-1)}/\epsilon^2)$ words of space. The above extends to all integers $p \geq 2$ if A is PSD.

The full proof is presented in Appendix A. We here sketch the proof. We take $G_1 \in \mathbb{R}^{1 \times n}$ and $G_2, G_3, \dots, G_p \in \mathbb{R}^{t \times n}$ as independent column normalized Gaussian matrix, where $t = O(n^{1-1/(p-1)})$. We then show an algorithm that computes in $\lceil p/2 \rceil$ -pass the estimator $X = G_1 A G_2^T G_2 \dots G_p A G_1^T$ and uses space at most t . As shown in Lemma 3.4, $X = \text{Tr}(X)$ is a unbiased estimator for Schatten p -norm with constant variance. By repeating the algorithm $O(1/\epsilon^2)$ times in parallel, we reach the desired accuracy.

3.4 Faster Update Time

Since Gaussian matrices are dense, a change to one coordinate of the input matrix A may lead to a change of every entry in the sketch. This means long update times for a streaming algorithm based on the sketch. In this section we extend our result for Gaussian sketching matrices to a distribution over $\{-1, 0, 1\}$ valued matrices with only one non-zero entry per column. The new sketch can be used to improve the update time of algorithms in the last two sections.

Definition 3.6 (Sparse ZD -sketch). *Let $\mathcal{D}_{t,n}$ be the distribution over matrices $G := ZD \in \mathbb{R}^{t \times n}$, where $Z = (z_1, z_2, \dots, z_n) \in \mathbb{R}^{t \times n}$ and $D = \text{diag}(d_1, d_2, \dots, d_n)$ are generated as follows. Let $h : [n] \rightarrow [t]$ be a 4-wise independent hash function, and set $Z_{i,j} = \mathbb{1}_{\{i=h(j)\}}$, i.e., in each z_j only the $h(j)$ -th coordinate is set to 1, and all other coordinates are 0. The diagonal entries of D are four-wise independent uniform $\{-1, 1\}$ random variables, and D is independent from Z .*

Notice that each column of G has a single non-zero entry, which is actually a random sign, and the n columns are four-wise independent. This random matrix G is similar to the sketching matrix used in [TZ04] to speed up the update time when estimating the second frequency moment of a vector in \mathbb{R}^n .

It is fairly easy to show that ZD -sketch works for approximating Schatten p -norm of matrices with all entries non-negative. The proof is shown in Appendix A.5. We now show that the conclusion of Theorem 3.2 and Theorem 3.5 still hold if replace the Gaussian matrices in the sketch with independent samples from the sparse ZD -sketch. A major difficulty that arises in replacing the Gaussian matrix with the sparse ZD -sketch is the latter's lack of rotational invariance. To prove Theorem 3.2 we were able to expand X^2 in terms of the eigenvalues of A and compute the expectation term-by-term, but this is not possible for the sparse ZD -sketch. For example, let G be a Gaussian matrix, for any orthogonal matrix U , the matrix GU is again a Gaussian matrix with an identical distribution to G . This does not hold for sparse ZD -sketch. As a consequence, in the expansion of $\mathbb{E}(X^2)$ in the proof of Proposition 3.3, the non-zero terms would also include those monomials of odd powers of $\lambda_i(A)$. For example, for Schatten 3-norm, one cannot bound $\sum_{i_1, i_2, \dots, i_6 \in [n]} \prod_{j=1}^6 \lambda_{i_j}$ by $O(\|A\|_{S_3}^6)$. But this term appears in the expansion of $\mathbb{E}(X^2)$ of the Schatten 3-norm estimator if using the sparse ZD -sketch matrices.

To resolve this problem, we use a technique similar to the proof of Hanson-Wright Inequality in [RV13] to bound the variance of X . The proof composed three major steps. The first step is to decouple the dependent summands by injecting independence. The second step is to replace the independent random vectors with fully independent Gaussian vectors while preserving the variance. We can then apply our techniques for Gaussians to bound the variance of the final random variable. The case $p = 1$ is useful to illustrate the technique, even though Schatten 1-norm approximation can be easily accomplished in other ways. Let $G \in \mathbb{R}^{t \times n}$ be the sparse JLT matrix and let $A \in \mathbb{R}^{n \times n}$ be PSD. The sketch is GAG^T and

$$\text{Tr}(GAG^T) - \text{Tr}(A) = \sum_{i \neq j} a_{i,j} \langle g_i, g_j \rangle. \quad (2)$$

Since $i \neq j$, g_i and g_j are independent. However the summands are subtly dependent. We first decouple the summand by choosing $\delta_i \sim \text{Bernoulli}(1/2)$, and write $\langle g_i, g_j \rangle = 4 \mathbb{E}(\delta_i(1 - \delta_j) \langle g_i, g_j \rangle)$. Let $V = \{i : \delta_i = 1\}$, then $\sum_{i \neq j} a_{i,j} \langle g_i, g_j \rangle = 4 \mathbb{E}_\delta \sum_{i \in V, j \in \bar{V}} a_{i,j} \langle g_i, g_j \rangle$. Thus conditioning on δ and $g_j : j \in \bar{V}$, $\langle g_i, \sum_{j \in \bar{V}} a_{i,j} g_j \rangle$ are independent random variables. We can match these random variables with Gaussian random variables of the same variance, and thus replace g_i with independent Gaussian vectors. The same process can be repeated for $g_j : j \in \bar{V}$, and replace every vector $g_i : i \in [n]$ by independent Gaussian vectors. This lets us apply similar techniques as used in the proof of Proposition 3.3 to bound the variance of the resulting random variable, and thus bound the variance of the original random variable $\text{Tr}(GAG^T) - \text{Tr}(A)$.

The analogue of (2) for the case of our general estimator, $X - \text{Tr}(A^p)$, is much more complicated than the $p = 1$ case. We observe that the terms of can be grouped as a sum of products of consecutive *walks*, i.e., $a_{i_1, i_2} a_{i_2, i_3} \dots a_{i_z, i_{z+1}} \langle g_{i_{z+1}}^{(z+1)}, g_{i_{z+1}}^{(z+1)} \rangle$. For each walk, we can apply similar idea to replace the g_i vectors with independent Gaussian vectors. Again, we apply similar techniques as used in the proof of Proposition 3.3 to bound the variance of each group. As a result, when replacing the Gaussian matrices by sparse JLT matrices, Lemma 3.4 still holds.

The proofs and technical details are presented in Section A.6. Using the sparse ZD -sketch, we are able to achieve the same space bound and query time as in Theorem 3.2 and Theorem 3.5. But our update time is improved to $O(1/\epsilon^2)$. The full and formal statement appears in Theorem A.11.

4 Unitarily Invariant Matrix Norms

In this section we show how to use a generic OSE to approximate any unitarily invariant matrix norm $l(A)$ using the sketch GAH^T for OSE matrices $G, H \in \mathbb{R}^{t \times n}$ and $t = (\text{intdim}(A)/\epsilon)^{O(1)}$. As we show in Appendix B, it is necessary that $t = (\text{sr}(A)/\epsilon)^{\Omega(1)}$, where $\text{sr}(A) := (\sum_{j \geq 1} \sigma_j^2)/\sigma_1^2 \leq \text{intdim}(A)$ is the *stable rank* of A . It is an open problem to determine whether a smaller sketch with $t = \text{sr}(A)^{O(1)}$ suffices for approximating $l(A)$. Theorem B.2 can be used to make some progress on the problem by showing that $t = \text{sr}(A)^{O(1)}$ dimensions suffices for a class of norms called Q -norms. OSEs embedding into $t = O(\text{sr}(A))$ are known to suffice for approximate matrix multiplication algorithms [CNW15].

Since we approximate $l(A)$ using a linear sketch, we immediately have a one-pass streaming algorithm. Furthermore, the sketch is bilinear and universal, i.e., it does not depend on the norm l except through the dimension of the sketch. Our main technical results describe the accuracy of the bilinear sketch for approximating functions of the input matrix. They are Theorem 4.1 for PSD matrices and Theorem B.2 for general matrices and Q -norms, which is postponed Appendix B.2. Their accuracy guarantees actually depend on the input's singular values, and thus their effectiveness depends on the norm at hand and/or on the intrinsic dimension (or stable rank) of the input.

Throughout, let I_d denote the identity matrix of order d . By convention, we extend a norm l on $n \times n$ matrices to smaller-order square matrices as follows. For a matrix $A \in \mathbb{R}^{d \times d}$ with $d < n$, we let $A' := \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}$ and define $l(A) := l(A')$.

Theorem 4.1. *Let $G \in \mathbb{R}^{t \times n}$ be an $(\epsilon, \delta, O(d \log n))$ -OSE matrix for integers $1 \leq d \leq t \leq n$ and some $\epsilon \in (1/n^9, 1/2)$ and $\delta \in (0, 1/2)$. Then for every PSD matrix $A \in \mathbb{R}^{n \times n}$, and every unitarily invariant norm $l(\cdot)$ on $n \times n$ matrices, with probability at least $1 - 2\delta n$,*

$$(1 - \epsilon)l(A) \leq l(GAG^T) \leq (1 + O(\epsilon))l(A) + O\left(\frac{l(I_d)}{d} \sum_{i=d+1}^n \lambda_i\right). \quad (3)$$

where $\lambda_1 \geq \dots \geq \lambda_n \geq 0$ are the eigenvalues of A .

We continue with applications of the theorem, before proving it in Section 4.1.

Low intrinsic dimension. The bound in (3) is particularly useful when the eigenvalues of A are decaying. Suppose we have an upper bound on the intrinsic dimension of A , namely, $\text{intdim}(A) \leq r$. Then we can bound $\sum_{i \geq 1} \lambda_i \leq r\lambda_1$, and by (3) we have that

$$(1 - \epsilon)l(A) \leq l(GAG^T) \leq (1 + O(\epsilon))l(A) + O\left(\frac{l(I_d)}{d} r\lambda_1\right). \quad (4)$$

As a first example, if $l(\cdot)$ is the Schatten p -norm for $p > 1$, we can choose $d = (r/\epsilon)^{1+1/(p-1)}$, which determines t^2 , the dimension of the sketch. We get that $(\|I_d\|_{S_p}/d) \cdot r\lambda_1 = d^{1/p-1}r\lambda_1 \leq \epsilon\|A\|_{S_p}$, and now (4) implies that $\|GAG^T\|_{S_p} \in (1 \pm O(\epsilon))\|A\|_{S_p}$.

A second example is the following straightforward extension. Fix $\alpha > 0$ and let l be a norm such that

$$\forall d \leq n, \quad l(I_d) \leq d^{1-\alpha}l(I_1).$$

We can choose $d = (r/\epsilon)^{1/\alpha}$, which determines t^2 , the dimension of the sketch. We get that $(l(I_d)/d) \cdot r\lambda_1 \leq d^{-\alpha}l(I_1) \cdot r\lambda_1 \leq \epsilon l(A)$, and now (4) implies that $l(GAG^T) \in (1 \pm O(\epsilon)) l(A)$.

A third example is where $l(\cdot)$ is the Ky Fan k -norm, defined as $\|A\|_{(k)} = \sigma_1 + \dots + \sigma_k$ where $\sigma_1 \geq \sigma_2 \geq \dots$ are the singular values of A . In our case where A is PSD, $\|A\|_{(k)} = \lambda_1 + \dots + \lambda_k$. By choosing $d = kr/\epsilon$, which determines t^2 , the dimension of the sketch, we get $(\|A\|_{(k)}/d) \cdot r\lambda_1 \leq (kr/d)\lambda_1 \leq \epsilon\|A\|_{(k)}$, and now (4) implies that

$$\|GAG^T\|_{(k)} \in (1 \pm O(\epsilon)) \|A\|_{(k)}.$$

Notice that all these examples use the same sketch, and more generally, a single sketch can be used to estimate many norms. In fact, with high probability the sketching succeeds — this is an event that does not depend on the desired norm — in which case the accuracy bounds (3) hold simultaneously for *all the relevant norms*. We note that this statement actually follows from our proof, although we do not formalize it for sake of brevity.

Proposition 4.2. *Fix $\alpha > 0$ and let \mathcal{L}_α be the set of all unitarily invariant norms l on $\mathbb{R}^{n \times n}$ such that $l(I_d) \leq d^{1-\alpha}l(I_1)$ for all $d \leq n$. There exists a randomized one-pass streaming algorithm \mathcal{A} that takes as input any PSD matrix $A \in \mathbb{R}^{n \times n}$ with $\text{intdim}(A) \leq r$, and outputs a square matrix D of order $O(r/\epsilon)^{1/\alpha}$, such that for every $l \in \mathcal{L}_\alpha$,*

$$\Pr \left[l(D) \in (1 \pm \epsilon) l(A) \right] \geq 9/10.$$

Moreover, \mathcal{A} uses $\text{poly}(r \log(n)/\epsilon)$ bits of space and processes each stream update in time $O(\epsilon^{-2} \log^2 n)$.

Proof. Algorithm \mathcal{A} follows Theorem 4.1 and maintains the linear sketch $D = GAG^T$. Its accuracy follows from the second example above, and it remains to bound the algorithm's space requirement and update time. If the OSE matrix G is constructed using Theorem 2.1, then every column of G has $O(\epsilon^{-1} \log n)$ non-zero entries, and an update to D (given an update to one entry of A) can be implemented in time $O(\epsilon^{-2} \log^2 n)$. The space used by the algorithm is only for storing the sketch matrix GAG^T , which has size $O(d^2) = O(r/\epsilon)^{2/\alpha}$ words. \square

4.1 Proof of Theorem 4.1

Before the proof we need three technical lemmas. Lemma 4.3 extends [AN13, Lemma 3.2] from (multiplying the input matrix by) Gaussian matrices to general OSEs. It asserts that the nonzero eigenvalues of SAS^T approximate that of A , including the sign. The proof of Theorem 4.1 is based on applying Lemma 4.3, separately for each “block” of eigenvalues (though A is not required to be block-diagonal).

Lemma 4.3. *Let $S \in \mathbb{R}^{t \times n}$ be an (ϵ, δ, d) -OSE and let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix of rank $r \leq d$. Then with probability at least $1 - \delta$, we have $\text{rank}(SAS^T) = r$ and there exists a permutation $\rho: [r] \rightarrow [r]$ such that*

$$\forall i \in [r], \quad \tilde{\lambda}_i(SAS^T) \in (1 \pm 3\epsilon) \tilde{\lambda}_{\rho(i)}(A),$$

where $\tilde{\lambda}_i(M)$ is the i th non-zero eigenvalue of M in decreasing order.

Lemma 4.4. Let $l(\cdot)$ be a unitarily invariant norm on matrices in $\mathbb{R}^{n \times n}$, and let $S \in \mathbb{R}^{t \times n}$ be an $(\epsilon, d\delta/n, d)$ -OSE matrix. Then for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, with probability at least $1 - \delta$,

$$l(SAS^T) \leq (1 + O(\epsilon)) \lceil n/d \rceil l(I_d) \|A\|_2,$$

where I_d is the identity matrix of order d .

Proposition 4.5. If $l(\cdot)$ is a symmetric norm on \mathbb{R}^m , then

$$\begin{aligned} l(A) &= \max l(|u_1^T Av_1|, |u_2^T Av_2|, \dots, |u_m^T Av_m|)^T \\ \text{s.t. } u_1, \dots, u_m &\in \mathcal{S}^{m-1} \text{ are orthogonal} \\ v_1, \dots, v_m &\in \mathcal{S}^{n-1} \text{ are orthogonal} \end{aligned}$$

We are now prepared to prove Theorem 4.1.

Proof of Theorem 4.1. We first show the second inequality in (3). By the Spectral Theorem, $A = U\Lambda U^T$, where $U \in \mathbb{R}^{n \times n}$ is orthonormal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. By Lemma 2.2, also $G' = GU$ is an $(\epsilon, \delta, O(d \log n))$ -OSE matrix, and we only need to consider the case where $A = \Lambda$ is a diagonal matrix. We group the eigenvalues into buckets. Let Λ_0 be the top d eigenvalues. For the rest, for $k = 1, 2, \dots$ define $\Lambda_k = \{\lambda_j : j > d \text{ and } \lambda_1/2^k < \lambda_j \leq \lambda_1/2^{k-1}\}$. Since the norm is monotonic in the eigenvalues of A , we can ignore eigenvalues smaller than $\epsilon\lambda_1/n$, i.e., treat them as zero. Indeed, by the triangle inequality, zeroing them can decrease $l(A)$ by at most $\epsilon\lambda_1 l(I_1) \leq \epsilon\ell(A)$. Thus, the effective number of buckets is $b = O(\log(n/\epsilon)) = O(\log n)$. For each $k \geq 1$, if bucket Λ_k is small, namely, $|\Lambda_k| \leq d$, then we add this bucket to Λ_0 (so now bucket Λ_k is empty). For simplicity, we use the same notation Λ_k for the new buckets (although formally this step defines a new set of buckets Λ'_k). With the new buckets, $|\Lambda_0| \leq (b+1)d = O(d \log n)$, and for every $k \geq 1$, either $|\Lambda_k| = 0$ or $|\Lambda_k| \geq d$.

With slight abuse of notation, we denote by Λ_k also the corresponding diagonal matrix, which is obtained from Λ by zeroing out the eigenvalues not in Λ_k . Thus $\Lambda = \sum_{k \geq 0} \Lambda_k$, and by the triangle inequality,

$$l(G\Lambda G^T) = l\left(\sum_{k \geq 0} G\Lambda_k G^T\right) \leq l(G\Lambda_0 G^T) + \sum_{k \geq 1} l(G\Lambda_k G^T).$$

By Lemma 4.3, with probability $1 - \delta$ we have $l(G\Lambda_0 G^T) \leq (1 + O(\epsilon)) l(\Lambda_0) \leq (1 + O(\epsilon))l(A)$. Let this event be denoted E_0 .

Now we apply Lemma 4.4 to each non-empty block $k = 1, \dots, b$. Specifically, in block k we apply the lemma treating l as a norm with dimension $|\Lambda_k|$. Thus, for each such k , with probability at least $1 - \delta|\Lambda_k|/d$, we have that $l(G\Lambda_k G^T) \leq (1 + O(\epsilon)) \frac{|\Lambda_k|}{d} l(I_d) \|\Lambda_k\|$. By a union bound over the b blocks, with probability at least $1 - \delta \frac{1}{d} \sum_{k \geq 1} |\Lambda_k| \geq 1 - \delta(n-d)/d$, we have

$$\sum_{k \geq 1} l(G\Lambda_k G^T) \leq (1 + O(\epsilon)) \frac{l(I_d)}{d} \sum_{k \geq 1} |\Lambda_k| \cdot \|\Lambda_k\|_2 \leq O(1) \frac{l(I_d)}{d} \sum_{i=d+1}^n \lambda_i.$$

Let this event be denoted E_1 . This completes the upper bound in (3).

Next, we show the lower bound on $\ell(GAG^T)$. The PSD matrices GAG^T and $\sqrt{\Lambda}G^TG\sqrt{\Lambda}$ have the same set of non-zero eigenvalues, hence $\ell(GAG^T) = \ell(\sqrt{\Lambda}G^TG\sqrt{\Lambda})$. By the variational characterization of $\ell(\cdot)$, as given in Proposition 4.5,

$$\ell(\sqrt{\Lambda}GG^T\sqrt{\Lambda}) \geq \ell(\text{diag}(e_1^T\sqrt{\Lambda}G^TG\sqrt{\Lambda}e_1, \dots, e_n^T\sqrt{\Lambda}G^TG\sqrt{\Lambda}e_n)).$$

For each $i \in [n]$, since G is an OSE matrix (for dimension at least 1), with probability at least $1 - \delta$ we have $e_i^T\sqrt{\Lambda}G^TG\sqrt{\Lambda}e_i = \lambda_i\langle Ge_i, Ge_i \rangle \geq (1 - \epsilon)\lambda_i$. By a union bound, with probability at least $1 - n\delta$ we have $\ell(GAG^T) \geq (1 - \epsilon)\ell(A)$, which event we denote by E_2 .

Combining the two inequalities we find with probability at least $1 - \Pr(\bar{E}_0) - \Pr(\bar{E}_1) - \Pr(\bar{E}_2) \geq 1 - 2n\delta$ both inequalities in (3) hold. \square

5 Top Singular Values

The results of [AN13] approximate the top eigen/singular values of a matrix using a bilinear sketch given by Gaussian matrices (note that if the input matrix is symmetric, then the signs of the eigenvalues can also be preserved). This section shows that similar guarantees can be achieved using a generic OSE instead of Gaussian matrix. By using a sparse OSE matrix, we obtain better update time than [AN13].

Theorem 5.1. *Fix $\epsilon > 0$ and integers $n, k \geq 1$. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix and let $\lambda_i^*(A)$ be the i -th largest eigenvalue of A in absolute value. Then there is a linear sketch of A , using space $O(\epsilon^{-2}k \log n)^2$, from which one can produce values $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_k$, such that*

$$\Pr \left[\forall i \in [k] : |\tilde{\lambda}_i(A) - \lambda_i^*(A)| \leq \epsilon |\lambda_i^*(A)| + \frac{1}{k} \sum_{i>k} |\lambda_i^*(A)| \right] \geq 5/9.$$

For each update to the matrix A , the sketch can be updated in time $O(\log^2 n / \epsilon^2)$.

Proof. The proof of this theorem follows from Lemma C.1 by setting $\phi = 1/(ck)$ for some constant $c \geq 2$. Indeed it follows that $\epsilon^2\phi \sum_{i=1/\phi+1} |\lambda_i^*(A)| \leq \frac{1}{k} \sum_{i=k+1} |\lambda_i^*(A)|$ and $|\lambda_{1/\phi}^*(A)| \leq \frac{1}{(c-1)k} \sum_{i=k+1} |\lambda_i^*(A)|$. The proof Lemma C.1 is very similar to the proof of Lemma 3.4 of [AN13], except that we give a more fine-grained statement regarding the mapping between the eigenvalues. Their sketch is of the form GAG^T , where G is a column-normalized Gaussian matrix. Our sketch is similar, except that G is an OSE matrix, specifically it is both $(\epsilon, 1/\text{poly } n, \phi^{-1})$ -OSE and $(0.1, 1/\text{poly } n, \phi^{-1}\epsilon^{-2})$ -OSE. The existence of such G follows from Theorem 2.1, and its sparsity determines our update time. \square

Theorem 5.2. *Fix $\epsilon > 0$ and integers $1 \leq m \leq n, k \geq 1$. Let $A \in \mathbb{R}^{n \times m}$ be a real matrix. Then there is a linear sketch of A , using space $O(k\epsilon^{-2} \log n)^2$, from which one can produce values $\hat{\sigma}_i$ for $i \in [k]$, satisfying the following with at least $5/9$ success probability,*

$$|\sigma_i^2(A) - \hat{\sigma}_i^2| \leq \epsilon \sigma_i^2(A) + \frac{1}{k} \sum_{i>k} \sigma_i^2(A).$$

For each update to the matrix A , the sketch can be updated in time $O(\log^2 n / \epsilon^2)$.

This theorem follows from the Lemma C.2 (by setting $\phi = O(1/k)$, with an identical argument as used in the proof of Theorem 5.1), which is a modification of [AN13, Lemma 3.5].

References

- [AC09] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009. doi:10.1137/060673096.
- [AKR15] A. Andoni, R. Krauthgamer, and I. Razenshteyn. Sketching and embedding are equivalent for norms. In *47th Annual ACM Symposium on Theory of Computing*, pages 479–488. ACM, 2015. doi:10.1145/2746539.2746552.
- [AL13] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013. doi:10.1145/2483699.2483701.
- [AN13] A. Andoni and H. Nguyễn. Eigenvalues of a matrix in the streaming model. In *24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1729–1737. SIAM, 2013. doi:10.1137/1.9781611973105.124.
- [BCKY15] V. Braverman, S. R. Chestnut, R. Krauthgamer, and L. F. Yang. Streaming symmetric norms via measure concentration. *CoRR*, abs/1511.01111, 2015. arXiv:1511.01111.
- [Bha97] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. doi:10.1007/978-1-4612-0653-8.
- [CNW15] M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015.
- [Coh16] M. B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 278–287. SIAM, 2016. doi:10.1137/1.9781611974331.ch21.
- [CW13] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *45th Annual ACM Symposium on Theory of Computing*, pages 81–90. ACM, 2013. doi:10.1145/2488608.2488620.
- [IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *30th Annual ACM Symposium on Theory of Computing*, pages 604–613. ACM, 1998. doi:10.1145/276698.276876.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984. doi:10.1090/conm/026.
- [JW13] T. S. Jayram and D. P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26, 2013. doi:10.1145/2483699.2483706.
- [KMN11] D. Kane, R. Meka, and J. Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 628–639. Springer, 2011. doi:10.1007/978-3-642-22935-0_53.
- [KN14] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4, 2014. doi:10.1145/2559902.
- [Le 14] F. Le Gall. Powers of tensors and fast matrix multiplication. In *39th International Symposium on Symbolic and Algebraic Computation*, pages 296–303. ACM, 2014. doi:10.1145/2608628.2608664.
- [Li11] Y. Li. Unpublished notes, 2011.
- [LNW14] Y. Li, H. L. Nguyen, and D. P. Woodruff. On sketching matrix norms and the top singular vector. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1562–1581. SIAM, 2014. doi:10.1137/1.9781611973402.114.

- [LW16a] Y. Li and D. P. Woodruff. On approximating functions of the singular values in a stream. In *48th Annual ACM Symposium on Theory of Computing*, pages 726–739. ACM, 2016. doi:10.1145/2897518.2897581.
- [LW16b] Y. Li and D. P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 60 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 39:1–39:11. Schloss Dagstuhl, 2016. doi:10.4230/LIPIcs.APPROX-RANDOM.2016.39.
- [NN13] J. Nelson and H. L. Nguyễn. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual Symposium on Foundations of Computer Science*, pages 117–126. IEEE, 2013. doi:10.1109/FOCS.2013.21.
- [RV13] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18(82):1–9, 2013. doi:10.1214/ECP.v18-2865.
- [Tro15] J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.
- [TZ04] M. Thorup and Y. Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, volume 4, pages 615–624, 2004.
- [Woo14] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10:1–157, 2014. doi:10.1561/04000000060.

A Proofs of Section 3

A.1 Proofs of Theorem 3.1

The proof of Theorem 3.1 relies on the following lemma.

Lemma A.1. *Let $G \in \mathbb{R}^{t \times n}$ be an $(\epsilon, \delta/n, 1)$ -JLT matrix. Then for every PSD matrix $A \in \mathbb{R}^{n \times n}$,*

$$\Pr \left[\text{Tr}(GAG^T) \in (1 \pm \epsilon) \text{Tr}(A) \right] \geq 1 - \delta.$$

Proof. By the Spectral Theorem, $A = U\Lambda U^T$, where Λ is a diagonal matrix and U is an orthonormal matrix. Then $G' = GU$ is a still $(\epsilon, \delta/n, 1)$ -JLT. Thus

$$\text{Tr}(GAG^T) = \text{Tr}(G'\Lambda G'^T) = \text{Tr}(\sqrt{\Lambda}G'^T G' \sqrt{\Lambda}) = \sum_{i=1}^n \lambda_i e_i^T G'^T G' e_i = \sum_{i=1}^n \lambda_i \|G' e_i\|_2^2.$$

By the JLT guarantee and a union bound, with probability at least $1 - \delta$, for all $i \in [n]$ we have $\|G' e_i\|_2^2 \in [1 - \epsilon, 1 + \epsilon]$, in which case $\text{Tr}(GAG^T) \in (1 \pm \epsilon) \text{Tr}(A)$. \square

Proof of Theorem 3.1. Apply Lemma A.1 to the PSD matrix $AA^T AA^T$, to get that with probability at least $1 - \delta$ (over the choice of G_2),

$$\text{Tr}(G_2 AA^T AA^T G_2^T) \in (1 \pm 2\epsilon) \text{Tr}(AA^T AA^T) = (1 \pm 2\epsilon) \|A\|_{S_4},$$

where we can rewrite the lefthand side as $\text{Tr}(AA^T G_2^T G_2 AA^T)$ using the identity $\text{Tr}(MM^T) = \text{Tr}(M^T M)$. Now suppose (by conditioning) that G_2 is already fixed, and apply the same lemma to the PSD matrix $AA^T G_2^T G_2 AA^T$, to get that with probability at least $1 - \delta$ (over the choice of G_1),

$$\text{Tr}(G_1 AA^T G_2^T G_2 AA^T G_1^T) \in (1 \pm 2\epsilon) \text{Tr}(AA^T G_2^T G_2 AA^T).$$

The proof follows by a union bound.

The linear sketch of A consists of the two matrices G_1A and G_2A , which suffices to estimate $\|A\|_{S_4}^4$ as above with $\delta = 1/8$. This sketch is linear and its dimension is $2tn$, where we can use say Gaussians to obtain $t = O(\epsilon^{-2} \log n)$. \square

A.2 Proofs of Proposition 3.3

Proof of Proposition 3.3. Using the identity $\text{Tr}(MM^T) = \text{Tr}(M^TM)$ we have $X = \text{Tr}(G_1AG_2^TG_2A \dots G_p^TG_pA \cdot G_1^T) = \text{Tr}(G_1^T \cdot G_1AG_2^TG_2A \dots G_p^TG_pA)$. By linearity of trace, expectation and matrix product, and by the fact that $\mathbb{E}[G_i^TG_i] = I_{n \times n}$ for all $i \in [p]$, we have

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\text{Tr}(G_1^T \cdot G_1AG_2^TG_2A \dots G_p^TG_pA)] \\ &= \mathbb{E}[\text{Tr}(I_{n \times n}AG_2^TG_2A \dots G_p^TG_pA)] \\ &= \dots = \text{Tr}(A^p). \end{aligned}$$

It remains to bound the variance of X . Without loss of generality we can assume that A is a diagonal matrix $\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, where $\lambda_1 \geq \dots \geq \lambda_n$. Indeed, in the case of a general symmetric A , we can write $A = U\Lambda U^T$ for an orthonormal matrix U and a diagonal matrix Λ . Then $G_iAG_{i+1}^T = (G_iU)\Lambda(G_{i+1}U)^T$, and the matrices $\{G_iU\}_{i \in [p]}$ have the same joint distribution as $\{G_i\}_{i \in [p]}$, hence $\text{Var}(X)$ would not change if A is replaced with Λ .

Let us write $G_i = (g_1^{(i)}, g_2^{(i)}, \dots, g_n^{(i)})$, where each $g_j^{(i)} \in \mathbb{R}^t$ is a column vector. It is easily verified that

$$X = \sum_{i_1, i_2, \dots, i_p \in [n]} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_p} \langle g_{i_1}^{(1)}, g_{i_2}^{(1)} \rangle \langle g_{i_2}^{(2)}, g_{i_3}^{(2)} \rangle \dots \langle g_{i_p}^{(p)}, g_{i_1}^{(p)} \rangle.$$

Indeed, first write $(G_1^TG_1A)_{i_1, i_2} = \sum_{k \in [t]} (G_1^T)_{i_1, k} (G_1)_{k, i_2} A_{i_2, i_2} = \langle g_{i_1}^{(1)}, g_{i_2}^{(1)} \rangle \lambda_{i_2}$, and then expand the trace in $X = \text{Tr}(G_1^TG_1A \cdot G_2^TG_2A \dots G_p^TG_pA)$ using all closed walks $(i_1, i_2, \dots, i_p) \in [n]^p$.

It is not difficult to verify that for all $j \neq j' \in [n]$ and $i_1, i_2, i'_1, i'_2 \in [p]$,

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)} \rangle] = \mathbb{1}_{\{i_1=i_2\}}, \quad (5)$$

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)} \rangle \langle g_{i'_1}^{(j')}, g_{i'_2}^{(j')} \rangle] = \mathbb{1}_{\{i_1=i_2, i'_1=i'_2\}}. \quad (6)$$

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)} \rangle \langle g_{i'_1}^{(j)}, g_{i'_2}^{(j)} \rangle] = \mathbb{1}_{\{(i_1, i'_1)=(i_2, i'_2)\}} + \frac{1}{t} \mathbb{1}_{\{(i_1, i_2)=(i'_1, i'_2)\}} + \frac{1}{t} \mathbb{1}_{\{(i_1, i_2)=(i'_2, i'_1)\}}. \quad (7)$$

Notice that in the last equation, the events in the three indicators are not disjoint, and when $i_1 = i'_1 = i_2 = i'_2$ the righthand-side evaluates to $1 + 2/t$.

We proceed to bound $\text{Var}(X) \leq \mathbb{E}[X^2]$. Denoting $I = (i_1, i_2, \dots, i_p) \in [n]^p$ with the convention $i_{p+1} := i_1$, and similarly for I' , we can write

$$X^2 = \left(\sum_I \prod_{j \in [p]} \lambda_{i_j} \langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)} \rangle \right)^2 = \sum_{I, I'} \prod_{j \in [p]} \lambda_{i_j} \lambda_{i'_j} \langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)} \rangle \langle g_{i'_j}^{(j)}, g_{i'_{j+1}}^{(j)} \rangle. \quad (8)$$

We can represent each term of X^2 (a fixed choice for I, I') by a diagram (see an example in Figure 1). Each node in the diagram represents an index i_j , and each square corresponds to a factor of the form $\langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)} \rangle \langle g_{i'_j}^{(j)}, g_{i'_{j+1}}^{(j)} \rangle$. A line connecting two nodes represents that the respective

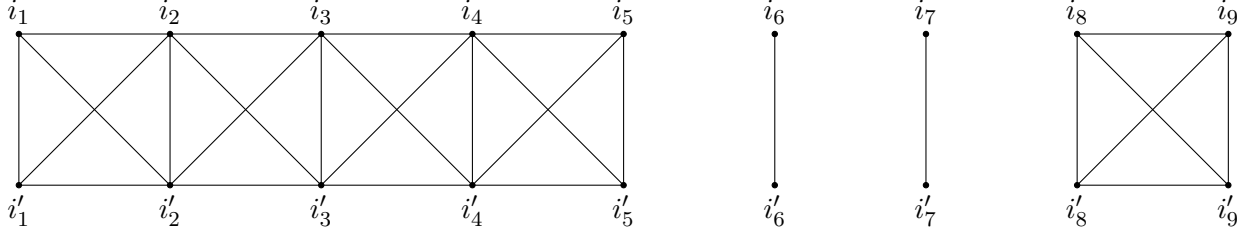


Figure 1: An example of a non-zero variance term ($p = 9$): $i_1 = \dots = i_5 = i'_1 = i'_2 = \dots = i'_5$, $i_6 = i'_6$, $i_7 = i'_7$, $i_8 = i_9 = i'_8 = i'_9$ and i_1, i_6, i_7, i_8 are distinct. The term of the eigenvalues in the variance expression is $\lambda_{i_1}^{10} \lambda_{i_6}^2 \lambda_{i_7}^2 \lambda_{i_8}^4$.

indices are equal. Notice that for each square, if a vertical line exists, then both vertical lines must exist, otherwise the expectation of this square is zero, and it has no contribution to $\mathbb{E}[X^2]$. Thus, for a non-zero diagram, if it has at least one vertical line, then it actually has all possible vertical lines. Diagrams with no vertical lines can be non-zero diagrams only if they are made entirely by horizontal cross-lines and parallel lines, which we call the *trivial diagrams*, and they correspond to the coefficient of $\lambda_i^p \lambda_j^p$ for $i \neq j$. Each non-trivial diagram corresponds to an *integer partition* of p (i.e., a way of writing the integer p as the sum of positive integers, with the order of the summands/parts having no significance), but we should account also for permutations and cyclic shifts on the parts. Given an integer partition $[p_1, p_2, \dots, p_z]$ of p , we write it as $(p_1^{(z_1)}, p_2^{(z_2)} \dots p_{t'}^{(z_{t'})})$, where $p_1 \geq p_2 \dots \geq p_{t'}$ are the distinct parts (or part sizes), and z_i counts how many parts are equal to p_i . Then the number of different diagrams for a given integer partition $[p_1, p_2, \dots, p_z]$ is

$$C_{[p_1, p_2, \dots, p_z]} = \frac{t'! p}{z_1! z_2! \dots z_{t'}!}.$$

Observe that this number is upper bounded by a constant M_p determined only by p . Each integer partition of p corresponds to a monomial of the eigenvalues. A connected component in the diagram corresponding to a power of the eigenvalue, and this power is just the size of that component. For each connected component, the total number of indices is an even number because of the vertical lines. For a single square, the coefficient is given by Equations (6)-(7). Using the diagram representation, we can calculate

$$\mathbb{E} \left(\begin{array}{|c|} \hline \square \\ \hline \end{array} \right) = 1 + \frac{2}{t}; \quad \mathbb{E} \left(\begin{array}{|c|} \hline \text{---} \\ \hline \text{---} \\ \hline \end{array} \right) = 1; \quad \mathbb{E} \left(\begin{array}{|c|} \hline | \\ \hline | \\ \hline \end{array} \right) = \mathbb{E} \left(\begin{array}{|c|} \hline \times \\ \hline \end{array} \right) = \frac{1}{t}. \quad (9)$$

All other diagrams either do not exist in the expansion of X^2 , or have a zero expectation. Diagrams corresponding to the same partition of p have the same coefficient. Since for each complete square there is a factor $1 + 2/t$, and for each incomplete square there is a factor $1/t$, the coefficient for a partition $[p_1, p_2, \dots, p_z]$ with $z > 1$ parts is

$$Z_{[p_1, p_2, \dots, p_z]} = \frac{1}{t^z} \left(1 + \frac{2}{t} \right)^{p-z}.$$

For non-trivial diagrams (i.e., have vertical lines) with $z > 1$ (i.e., excluding the completely connected graph) we collect all such terms as X_1 and bound their expectation by

$$\mathbb{E}[X_1] \leq \sum_{[p_1, p_2, \dots, p_z]} \frac{M_p}{t^z} \sum_{i_1, i_2, \dots, i_z \in [n]} \lambda_{i_1}^{2p_1} \lambda_{i_2}^{2p_2} \dots \lambda_{i_z}^{2p_z}. \quad (10)$$

For non-trivial diagrams and $z = 1$, there cannot be any incomplete square, and we can compute the expression explicitly,

$$\mathbb{E} \left[\sum_{i \in [n]} \lambda_i^{2p} \prod_{j \in [p]} \langle g_i^{(j)}, g_i^{(j)} \rangle \langle g_i^{(j)}, g_i^{(j)} \rangle \right] = \left(1 + \frac{2}{t}\right)^p \sum_{i=1}^n \lambda_i^{2p} = 3^p \|A\|_{S_{2p}}^{2p}.$$

For trivial diagrams (no vertical lines), we collect the terms as X_2 and bound their expectation by

$$\mathbb{E}[X_2] \leq \sum_{i \neq k \in [n]} \lambda_i^p \lambda_k^p \sum_{z=0}^p \binom{p}{z} \mathbb{E} \left(\bigotimes^z \right) \mathbb{E} \left(\bigotimes^{p-z} \right) \leq 2^p \sum_{i \neq k \in [n]} \lambda_i^p \lambda_k^p \leq M'_p \|A\|_{S_p}^{2p}, \quad (11)$$

where M'_p is a constant that depends only on p .

We now turn to bounding $\mathbb{E}[X_1]$ using (10). For each integer partition $[p_1, p_2, \dots, p_z]$ of p with $z > 1$ parts,

$$\sum_{i_1, i_2, \dots, i_z \in [n]} \lambda_{i_1}^{2p_1} \lambda_{i_2}^{2p_2} \dots \lambda_{i_z}^{2p_z} = \left(\sum_{i_1 \in [n]} \lambda_{i_1}^{2p_1} \right) \dots \left(\sum_{i_z \in [n]} \lambda_{i_z}^{2p_z} \right) = \prod_{j=1}^z \|A\|_{S_{2p_j}}^{2p_j}.$$

Let z' be the number of parts with $2p_j \leq p$. Clearly, $z' \geq z - 1$, since at most one part can have $p_j \geq p/2$. Consider first the case $z' = z$. It is well-known (via an application of Hölder's inequality) that $\|x\|_q \leq \|x\|_r \leq n^{1/r-1/q} \|x\|_q$ holds for all $x \in \mathbb{R}^n$ and $1 \leq r \leq q$. This comparison of norms applies also to the Schatten norms of A (viewed as n -dimensional norms of the eigenvalues of A), proves that $\|A\|_{S_{2p_j}}^{2p_j} \leq n^{1/(2p_j)-1/p} \|A\|_{S_p}^{2p_j}$. We thus obtain

$$\prod_{j=1}^z \|A\|_{S_{2p_j}}^{2p_j} \leq \prod_{j=1}^z \left(n^{1/(2p_j)-1/p} \|A\|_{S_p}^{2p_j} \right)^{2p_j} = n^{z-2} \|A\|_{S_p}^{2p}. \quad (12)$$

In the case $z' = z - 1$, there is a unique j^* such that $p_{j^*} > p/2$, and therefore $z \leq (p - p_{j^*}) + 1 \leq \lfloor p/2 \rfloor + 1$. For $j \neq j^*$ we can use the comparison of Schatten norms as above, and for $j = j^*$ we simply use $\|A\|_{S_{2p_j}} \leq \|A\|_{S_{2p}}$. We thus obtain

$$\prod_{j=1}^z \|A\|_{S_{2p_j}}^{2p_j} \leq \|A\|_{S_{2p}}^{2p_{j^*}} \prod_{j \neq j^*} \left(n^{1/(2p_j)-1/p} \|A\|_{S_p}^{2p_j} \right)^{2p_j} \leq n^{z-1-2(p-p_{j^*})/p} \|A\|_{S_p}^{2p} \leq n^{z-2z/p} \|A\|_{S_p}^{2p}. \quad (13)$$

where the last inequality follow by $1 + 2(p - p_{j^*})/p \geq 1 + 2(z - 1)/p = 2z/p + (1 - 2/p)$. With also the $z = 1$ term considered, we have

$$\text{Var}(X) \leq M''_p \left(1 + \sum_{z=2}^{\lfloor p/2 \rfloor + 1} \left(\frac{n^{1-2/p}}{t} \right)^z + \sum_{z=2}^p \left(\frac{n^{1-2/z}}{t} \right)^z \right) \|A\|_{S_p}^{2p}.$$

where M''_p is a constant depends only on p . This completes the proof of Proposition 3.3. \square

A.3 Proofs of Lemma 3.4

Proof of Lemma 3.4. We first argue that it suffices to prove the corollary under the assumption that the entries of G_l are fully independent. Indeed, each of the terms we need to calculate is the expectation of a polynomial of total degree at most 4 in the random variables G_{ij} . For example, the factor contains G_l in a typical term of X^2 is $\langle g_{i_l}^{(l)} g_{j_l}^{(l)} \rangle \langle g_{i_l'}^{(l)} g_{j_l'}^{(l)} \rangle$. The expectation of such a polynomial when G_l 's entries are 4-wise independent is exactly the same as when these entries are fully independent.

Assume henceforth that the entries of G_l are mutually independent. We repeat the proof of Proposition 3.3, except that when considering the square containing (i_1, i_2) , we replace t with t' in (7) and (9). In diagrams where this square is complete, the contribution to $\mathbb{E}[X^2]$, as given by (8), does not change. When this square is incomplete, we replace t by t' in subsequent calculations like (10) and (11). The proof is otherwise identical, but we kept the more precise bound obtained in (13). \square

A.4 Proofs of Theorem 3.5

Proof of Theorem 3.5. Without loss of generality we may assume that A is symmetric as argued in the proof of Theorem 3.2. We first describe a basic algorithm that produces an estimator for $\|A\|_{S_p}^{2p}$ that is unbiased and has variance $O_p(\|A\|_{S_p}^{2p})$. We will later decrease the variance to $O(\epsilon^2 \|A\|_{S_p}^{2p})$ using the standard technique of independent repetitions in parallel.

The basic algorithm uses a pseudo-random generator to produce a four-wise independent column-normalized Gaussian matrix. In fact, it samples p such matrices, namely, $G_1 \in \mathbb{R}^{1 \times n}$ and $G_2, \dots, G_p \in \mathbb{R}^{t \times n}$ for $t = O(n^{1-1/(p-1)})$, where the p matrices are independent of each other. In the first pass, the algorithm computes $G_1 A G_2^T$ and $G_p A G_1^T$, and stores them in memory. Notice that these are linear sketches of A , each dimension t . In the second pass, the algorithm uses these results to compute $(G_1 A G_2^T) G_2 A G_3^T$ and $G_{p-1} A G_p^T (G_p A G_1^T)$ which are again linear sketches of the stream A (given the result of the first pass), each of dimension t . Continuing in this manner until pass number $\lceil p/2 \rceil$, the algorithm stores in memory the vectors $h = G_{\lceil p/2 \rceil} A G_{\lceil p/2 \rceil+1}^T \dots G_p A G_1^T$ and $h^T = G_1 A G_2^T \dots G_{\lceil p/2 \rceil-1} A G_{\lceil p/2 \rceil}^T$, each of dimension t . Now compute $Y = h^T h$. By Lemma 3.4 we have that $\mathbb{E}[Y] = \text{Tr}(A^p) = \sum_i \lambda_i^p$ (where in the case that p is odd we use the assumption that A is PSD). Thus, Y is an unbiased estimator for $\|A\|_{S_p}^{2p}$, and it remains to bound its variance. By Lemma 3.4,

$$\text{Var}(Y) = O_p \left(\sum_{z=2}^{\lceil p/2 \rceil} \frac{n^{z-1-2(z-1)/p}}{n^{z-1-(z-1)/(p-1)}} + \sum_{z=2}^p \frac{n^{z-2}}{n^{z-1-(z-1)/(p-1)}} \right) \|A\|_{S_p}^{2p} = O_p(\|A\|_{S_p}^{2p}).$$

By repeating the basic algorithm $O_p(1/\epsilon^2)$ times in parallel and reporting the average of their estimates Y , we obtain estimator X for $\|A\|_{S_p}^{2p}$ that is unbiased and has variance at most $\frac{1}{9}\epsilon^2 \|A\|_{S_p}^{2p}$. The correctness of this estimator follows by Chebyshev's inequality. The basic algorithm is required to store $2p$ intermediate vectors of dimension t and random seeds for the p Gaussian matrices. By standard techniques, the length of the seeds is $O_p(\text{polylog } n)$ bits. The final algorithm stores these for all the $O_p(1/\epsilon^2)$ repetitions, and Theorem 3.5 follows. \square

A.5 A Simple Proof for Sparse Sketch of Matrices With Non-Negative Entries

Lemma A.2. Let $G = (g_1, g_2, \dots, g_n) \sim \mathcal{D}_{t,n}$, then the following satisfies,

1. for each $i \in [n]$ $E(\langle g_i, g_i \rangle) = 1$, $E(\langle g_i, g_i \rangle^2) = 1$;
2. for each $i, j \in [n], i \neq j$ $E(\langle g_i, g_j \rangle) = 0$, $E(\langle g_i, g_j \rangle^2) = 1/t$;
3. for each $i, j, i', j' \in [n], \{i, j\} \neq \{i', j'\}, i \neq j, i' \neq j'$ $E(\langle g_i, g_j \rangle) = 0$, $E(\langle g_i, g_j \rangle \langle g_{i'}, g_{j'} \rangle) = 0$;

Proof. Property 1 follows immediately. For 2, $E(\langle g_i, g_j \rangle) = 0$ and

$$\begin{aligned} E(\langle g_i, g_j \rangle^2) &= E\left(\sum_l g_{i,l} g_{j,l}\right)^2 \\ &= \sum_{l,k} E(g_{i,l} g_{j,l} g_{i,k} g_{j,k}) \\ &= \sum_{l=1}^t E(d_i^2 d_j^2 z_{i,l} z_{i,k}) \\ &= \sum_{l=1}^t \frac{1}{t^2} = \frac{1}{t}. \end{aligned}$$

For 3, we only need to consider the case when $\{i, j\} \cap \{i', j'\} \neq \emptyset$. Without loss of generality, assume $i = i'$, thus,

$$E(\langle g_i, g_j \rangle \langle g_i, g_{j'} \rangle) = \sum_l E(g_{i,l} g_{j,l} g_{i,l} g_{j',l}) + \sum_{l \neq k} E(g_{i,l} g_{j,l} g_{i,k} g_{j',k}) = 0,$$

where we use that $g_{i,l} g_{i,k} = 0$ when $l \neq k$. □

The following lemma is a simple case that the variance of sparse sketch is smaller than the Gaussian sketch. We will show in the next section that the sparse sketch is superior than Gaussian sketch for every symmetric matrix.

Lemma A.3. Let $G_1 \sim \mathcal{D}_{t',n}$ be and G_2, \dots, G_p be independent copy of $\mathcal{D}_{t,n}$, where $p \geq 2$ is a integer and c_1, c_2 are two absolute constants. Let A be a symmetric matrix with all entries non-negative and $1 \leq t' \leq t$. Let $X = \text{Tr}(G_1 A G_2^T G_2 A G_3 \dots G_p A G_1^T)$. Let X' be a random variable by replacing G_i of X by column normalized gaussian matrix of the same size. Then,

$$E(X^2) \leq E(X'^2).$$

Proof. Let $J = \{j_1, j_2, \dots, j_p\} \in [n]^p$ $I = \{i_1, i_2, \dots, i_p\} \in [n]^p$. Define

$$X_{I,J} := a_{i_p, j_1} a_{i_1, j_2} \dots a_{i_{p-1}, j_p} \langle g_{j_1}^{(1)}, g_{i_1}^{(1)} \rangle \langle g_{j_2}^{(2)}, g_{i_2}^{(2)} \rangle \dots \langle g_{j_p}^{(p)}, g_{i_p}^{(p)} \rangle.$$

We now expand X in a different form,

$$X = \sum_{I,J} X_{I,J}.$$

Thus,

$$X^2 = \sum_{I,J,I',J'} X_{I,J} X_{I',J'}.$$

Define $X'_{I,J}$ analogously by replacing g_i with gaussian vectors. Since each $a_{i,j} \geq 0$, with Proposition 3.3 and Lemma A.2, we immediately have that $E(X^2) \leq E(X'^2)$. \square

The above lemma leads to the following theorem,

Theorem A.4. *For every integer $p \geq 2$, there exist a randomized one-pass streaming algorithm \mathcal{A} using space $O(n^{2-4/p}/\epsilon^2)$, and a $\lceil p/2 \rceil$ -pass streaming algorithm \mathcal{B} using space $O(n^{1-1/(p-1)}/\epsilon^2)$, given as input PSD matrix $A \in \mathbb{R}^{n \times n}$ with all entries non-negative, then the output of the algorithms $\mathcal{A}(A)$ and $\mathcal{B}(A)$ satisfy*

$$\Pr[\mathcal{A}(A) \in (1 \pm \epsilon)\|A\|_{S_p}^p] \geq 0.99;$$

and

$$\Pr[\mathcal{B}(A) \in (1 \pm \epsilon)\|A\|_{S_p}^p] \geq 0.99,$$

where the probability is over the randomness of the algorithms. Both algorithms require $O(1/\epsilon^2)$ time to process each update in a pass. After the updates, \mathcal{A} requires time $O(n^{(1-2/p)\omega}/\epsilon^2)$ to compute its output and \mathcal{B} requires time $O(n^{(1-2/p)}/\epsilon^2)$, where $\omega < 3$ is the matrix multiplication constant. For general matrix input A of size $n \times m$ for $m \leq n$, if A has all entries non-negative, the above claim holds for even integer $p \geq 2$.

A.6 Proofs of Sparse Sketch for General Matrices

Definition A.5 (Bounded Variance). *We say that a random vector $g \in \mathbb{R}^t$ has bounded variance if,*

$$\sup_{u \in S^{t-1}} \mathbb{E}(\langle g, u \rangle^2) \leq C,$$

for some absolute constant $C > 0$.

Proposition A.6 (Rotational Invariance of Centered Random Variables). *Let X_1, X_2, \dots, X_n be independent centered random variables with mean 0 and variance at most $K > 0$. Then for every $a = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$,*

$$\mathbb{E} \left[\left(\sum_{i=1}^n a_i X_i \right)^2 \right] \leq K \|a\|_2^2.$$

In particular, the random vector $(X_1, X_2, \dots, X_n)/K \in \mathbb{R}^n$ has bounded variance.

Lemma A.7. *Let $H_1, H'_1, H_2, H'_2, \dots, H_p, H'_p \in \mathbb{R}^{t \times n}$ be independent column-normalized Gaussian matrices. Then for every matrices $A_1, A_2, \dots, A_p \in \mathbb{R}^{n \times n}$, the random variable $X = \text{Tr}(H'_1 A_1 H_2^T H'_2 A_2 H_3^T H'_3 A_3 \dots H'_p A_p H_1^T)$ satisfies*

$$\mathbb{E}(X^2) \leq \frac{1}{t^p} \prod_i \|A_i\|_F^2.$$

Proof. Let $A_i = U_i \Sigma_i V_i$ be the singular value decomposition of A_i . Then X is identically distributed as $X' = \text{Tr}(\prod_i H_i^T H'_i \Sigma_i)$. Now apply the same technique as used in the proof Proposition 3.3, and the lemma follows. \square

Lemma A.8. Let $A_1, A_2, \dots, A_r \in \mathbb{R}^{n \times n}$ be arbitrary matrices for an integer $r \geq 1$. Let $G_1, G_2, \dots, G_r \in \mathbb{R}^{t \times n}$ be independent random matrices such that for each $l \in [r]$, $\sqrt{t} \cdot G_l$ has i.i.d. columns with bounded variance. Denote $A_i = \left(a_{j,k}^{(i)} \right)_{\{j,k \in [n]\}}$, and $G_i = \left(g_1^{(i)}, g_2^{(i)}, \dots, g_n^{(i)} \right)$. Let

$$X := \sum_{\substack{i_1 \neq j_1 \in [n], \\ \dots, \\ i_r \neq j_r \in [n]}} \prod_{l=1}^r \left\langle g_{i_l}^{(l)}, g_{j_l}^{(l)} \right\rangle a_{j_l, i_{l+1}}^{(l)},$$

where by convention $i_{r+1} = i_1$. Then,

$$\mathbb{E}[X^2] \leq \frac{C^r}{t^r} \prod_{i=1}^r \|A_i\|_F^2,$$

for some absolute constant $C > 0$.

Proof of Lemma A.8. We use a similar proof structure as in [RV13].

Decoupling. Let $\delta_i^{(l)} \sim U\{0, 1\}$ be independent Bernoulli random variables with $\mathbb{E}(\delta_i^{(l)}) = 1/2$. We have that

$$X = 4^r \mathbb{E}_\delta \left[\sum_{\substack{i_1 \neq j_1 \in [n], \\ \dots, \\ i_r \neq j_r \in [n]}} \prod_{l=1}^r \delta_{i_l}^{(l)} (1 - \delta_{j_l}^{(l)}) \left\langle g_{i_l}^{(l)}, g_{j_l}^{(l)} \right\rangle a_{j_l, i_{l+1}}^{(l)} \right].$$

Let $V_l = \{i \in [n] : \delta_i^{(l)} = 1\}$. Then setting

$$X_\delta := \sum_{\substack{i_1 \in V_1, j_1 \in \bar{V}_1, \\ \dots, \\ i_r \in V_r, j_r \in \bar{V}_r}} \prod_{l=1}^r \left\langle g_{i_l}^{(l)}, g_{j_l}^{(l)} \right\rangle a_{j_l, i_{l+1}}^{(l)}.$$

we can write $X = 4^r \mathbb{E}_\delta[X_\delta]$. Now by Jensen's inequality,

$$\mathbb{E}_G[X^2] = 8^r \mathbb{E}_G[(\mathbb{E}_\delta[X_\delta])^2] \leq 8^r \mathbb{E}_{\delta, G}[X_\delta^2].$$

Rewrite X_δ as

$$X_\delta = \frac{1}{t} \sum_{i_1 \in V_1} \left\langle \sqrt{t} g_{i_1}^{(1)}, \sqrt{t} \sum_{j_1 \in \bar{V}_1} M_{i_1, j_1}^1 \cdot g_{j_1}^{(1)} \right\rangle,$$

where

$$M_{i_1, j_1}^1 = \sum_{\substack{i_2 \in V_2, j_2 \in \bar{V}_2, \\ \dots, \\ i_r \in V_r, j_r \in \bar{V}_r}} \prod_{l=2}^r \left\langle g_{i_l}^{(l)}, g_{j_l}^{(l)} \right\rangle a_{j_l, i_{l+1}}^{(l)}.$$

Since $\sqrt{t}g_{i_1} \in \mathbb{R}^t$ is centered with bounded variance, after conditioning on V_1 and on $\{g_j : j \in \bar{V}_1\}$ (and viewing $M_{i,j}^1$ as fixed for now), for each $i \in V_1$ we have that

$$\left\langle \sqrt{t}g_i^{(1)}, \sum_{j \in \bar{V}_1} M_{i,j}^1 \sqrt{t}g_j^{(1)} \right\rangle / \left\| \sum_{j \in \bar{V}_1} M_{i,j}^1 \sqrt{t}g_j^{(1)} \right\|_2$$

is an independent centered random variable with bounded variance. By rotational invariance of centered random variables (that have bounded variance), we have that

$$\mathbb{E}_{\{g_i^{(1)} : i \in V_1\}} \left[X_\delta^2 \mid \delta, g_j^{(1)} : j \in \bar{V}_1 \right] \leq \frac{c_0}{t^2} \sum_{i \in V_1} \left\| \sum_{j \in \bar{V}_1} M_{i,j}^1 \sqrt{t}g_j^{(1)} \right\|_2^2.$$

Reduction. In this step, we replace the random vectors by Gaussian vectors. Let

$$Z := \frac{1}{t} \sum_{i \in V_1} \left\langle \sqrt{t}h_i^{(1)}, \sum_{j \in \bar{V}_1} M_{i,j}^1 \sqrt{t}g_j^{(1)} \right\rangle,$$

where $h_i^{(1)}$ has independent entries from $N(0, 1/t)$. Then

$$\mathbb{E}_H(Z^2) = \frac{1}{t^2} \sum_{i \in V_1} \left\| \sqrt{t} \sum_{j \in \bar{V}_1} M_{i,j}^1 g_j^{(1)} \right\|_2^2.$$

Thus $\mathbb{E}_{g_{V_1}^{(1)}}(X_\delta^2) \leq \mathbb{E}_H(c_1 Z^2)$, i.e., we have replaced some of the g vectors with gaussian vectors h .

We now rewrite

$$Z = \frac{1}{t} \sum_{j \in \bar{V}_1} \left\langle \sqrt{t}g_j^{(1)}, \sum_{i \in V_\delta} M_{i,j}^1 \sqrt{t}h_i^{(1)} \right\rangle,$$

and repeat the above process for $g_j^{(1)}, g_{i_2}^{(2)}, g_{j_2}^{(2)}, \dots, g_{i_r}^{(r)}, g_{j_r}^{(r)}$, we reach

$$\mathbb{E}_{G,\delta}[X_\delta^2] \leq c_2^r \mathbb{E}_H[Y_\delta^2],$$

where

$$Y_\delta := \sum_{\substack{i_1 \in V_1, j_1 \in \bar{V}_1, \\ \dots, \\ i_r \in V_r, j_r \in \bar{V}_r}} \prod_{l=1}^r \left\langle h_{i_l}^{(l)}, h_{j_l}^{(l)} \right\rangle a_{j_l, i_{l+1}}^{(l)} = \text{Tr}(H'_1 A'_1 H_2^T H'_2 A'_2 H_3^T H'_3 A'_3 \dots H'_r A'_r H_1^T),$$

where $A'_l = P_{\delta(l)} A_l (I - P_{\delta(l)})$, and P_δ is the restriction projection of \mathbb{R}^n into \mathbb{R}^{V_δ} .

Calculation Using Gaussian Random Variables. By Lemma A.7,

$$\mathbb{E}_H[Y_\delta^2] \leq \frac{1}{t^r} \prod_{i=1}^r \|A'_i\|_F^2 \leq \frac{1}{t^r} \prod_{i=1}^r \|A_i\|_F^2.$$

Notice that the righthand-side does not depend on δ , and we thus conclude that

$$\mathbb{E}[X^2] \leq c_3^r \mathbb{E}_\delta[\mathbb{E}_H[Y_\delta^2]] \leq c_4^r \frac{1}{t^r} \prod_{i=1}^r \|A_i\|_F^2.$$

□

Proposition A.9. *Let $g \in \mathbb{R}^t$ be a column of $G \sim \mathcal{D}_{t,n}$. Then $\sqrt{t}g$ is a centered random vector with bounded variance.*

Proof. Let $u \in S^{t-1}$ be any unit vector. Let

$$X := \langle \sqrt{t}g, u \rangle = \sum_{i=1}^t \sqrt{t}g_i u_i.$$

Thus $E(X) = 0$ and by expanding X^2 we get that $E(X^2) = \sum_i u_i^2 = 1$. □

Lemma A.10. *Let $A \in \mathbb{R}^{n \times n}$ be an arbitrary symmetric matrix. Let G_1, G_2, \dots, G_p be independent copies of $\mathcal{D}_{t,n}$. Denote $A = (a_{i,j})_{\{i,j \in [n]\}}$ and $G_i = (g_1^{(i)}, g_2^{(i)}, \dots, g_n^{(i)})$. Then*

$$X := \text{Tr}(G_1 A G_2^T G_2 A G_3^T \dots G_p A G_1^T),$$

satisfies

$$\mathbb{E}(X) = \text{Tr}(A^p)$$

and

$$\text{Var}(X) = O_p \left(1 + \sum_{z=2}^{\lfloor p/2 \rfloor + 1} \left(\frac{n^{1-2/p}}{t} \right)^z + \sum_{z=2}^p \left(\frac{n^{1-2/z}}{t} \right)^z \right) \|A\|_{S_p}^{2p}.$$

Proof. Observe that $X = \text{Tr}(G_1 A G_2^T G_2 A G_3^T \dots G_p A G_1^T) = \text{Tr}(A G_1^T G_1 \dots A G_p^T G_p)$ and write

$$(A G_1^T G_1 \dots A G_p^T G_p)_{i_p, i_p} = \sum_{i_1, \dots, i_{p-1} \in [n]} \prod_{l=1}^p (A G_l^T G_l)_{i_{l-1}, i_l} = \sum_{\substack{i_1, \dots, i_{p-1} \in [n], \\ j_1, \dots, j_p \in [n]}} \prod_{l=1}^p a_{i_{l-1}, j_l} (G_l^T G_l)_{j_l, i_l},$$

where $(G_l^T G_l)_{j_l, i_l} = \langle g_{j_l}^{(l)}, g_{i_l}^{(l)} \rangle$. We can thus expand X as

$$X = \sum_{I, J \in [n]^p} X_{I, J},$$

where for each $J = (j_1, j_2, \dots, j_p) \in [n]^p$ and $I = (i_1, i_2, \dots, i_p) \in [n]^p$, we define

$$X_{I, J} := a_{i_p, j_1} a_{i_1, j_2} \dots a_{i_{p-1}, j_p} \langle g_{j_1}^{(1)}, g_{i_1}^{(1)} \rangle \langle g_{j_2}^{(2)}, g_{i_2}^{(2)} \rangle \dots \langle g_{j_p}^{(p)}, g_{i_p}^{(p)} \rangle.$$

For a given pair (I, J) , if some $i_l \neq j_l$ then $\mathbb{E}(X_{I, J}) \propto \mathbb{E}(\langle g_{j_l}^{(l)}, g_{i_l}^{(l)} \rangle) = 0$, and now the independence between the different matrices G_l implies

$$\mathbb{E}[X] = \sum_{i_1, \dots, i_p \in [n]} a_{i_1, i_2} a_{i_2, i_3} \dots a_{i_p, i_1} = \text{Tr}(A^p).$$

Let us say that a pair (I, J) is a *cycle* if $i_1 = j_1, i_2 = j_2, \dots, i_p = j_p$. We partition the other pairs (i.e., non-cycle terms $X_{I,J}$) as follows. A sequence of indices $i_k, j_k, i_{k+1}, j_{k+1}, \dots, j_{k+z-1}$ defines a *walk* if $i_k = j_k, i_{k+1} = j_{k+1}, \dots, i_{k+z-1} = j_{k+z-1}$ and $i_{k-1} \neq j_{k-1}, i_{k+z} \neq j_{k+z}$ (where subscripts are viewed modulo p , e.g. $j_{p+1} = j_1$). This walk corresponds to a factor

$$W(k, z) := a_{i_{k-1}, i_k} a_{i_k, i_{k+1}} \dots a_{i_{k+z-2}, i_{k+z-1}} a_{i_{k+z-1}, j_{k+z}} \left\langle g_{j_{k+z}}^{(k+z)}, g_{i_{k+z}}^{(k+z)} \right\rangle,$$

where we omitted “1-factors” of the form $\langle g_{i_l}^{(l)}, g_{i_l}^{(l)} \rangle = 1$. Such a walk W can be defined by two integers k and z , representing that the walk starts from i_{k-1} and has length z . Let us say that $W(k_2, z_2)$ is *consecutive* to $W(k_1, z_1)$ if $k_1 + z_1 = k_2$. We can now represent each non-cycle term $X_{I,J}$ as a product of consecutive walks (recall we omit “1-factors”).

Let $z = [z_1, z_2, \dots, z_r]$ be an ordered integer partition of p , i.e., $z_1 + z_2 + \dots + z_r = p$ and all $z_l \geq 1$. The sum of all $X_{I,J}$ whose walks match the partition $z = [z_1, z_2, \dots, z_r]$ can be written as

$$T_z = \frac{1}{R(z)} \sum_{k \in [p]} \sum_{\substack{I, J: \\ \text{walks satisfy } k \text{ and } z}} W(k, z_1) \cdot W(k + z_1, z_2) \cdot \dots \cdot W(k + z_1 + z_2 + \dots + z_{r-1}, z_r),$$

where $R(z)$ counts how many different starting point $k \in [p]$ yield the same walk, thus $R(z) \leq p$.

$$X - \text{Tr}(A^p) = \sum_{\text{partition } z} T_z. \quad (14)$$

For each z , we can write, using a cyclic shift of the indices (to map k to 1),

$$\begin{aligned} T_z = C_z \sum_{\substack{i_1 \neq j_1, \\ i_{z_1+1} \neq j_{z_1+1}, \\ \vdots \\ i_{z_1+z_2+\dots+z_{r-1}} \neq j_{z_1+z_2+\dots+z_{r-1}}, \\ \vdots \\ i_{z_1+z_2+\dots+z_{r-1}+1} \neq j_1}} b_{i_1, j_{z_1+1}} \langle g_{j_{z_1+1}}^{(z_1+1)}, g_{i_{z_1+1}}^{(z_1+1)} \rangle \cdot b_{i_{z_1+1}, j_{z_1+z_2+1}} \langle g_{j_{z_1+z_2+1}}^{(z_1+z_2+1)}, g_{i_{z_1+z_2+1}}^{(z_1+z_2+1)} \rangle \\ \cdot \dots \cdot b_{i_{z_1+z_2+\dots+z_{r-1}+1}, j_1} \langle g_{j_1}^{(1)}, g_{i_1}^{(1)} \rangle \end{aligned} \quad (15)$$

where C_z is a constant that depends only on z and

$$b_{i_r, j_l} := \sum_{i_{r+1}, \dots, i_{l-1}} a_{i_r, i_{r+1}} a_{i_{r+1}, i_{r+2}} \dots a_{i_{l-1}, j_l} = (A^{l-r})_{i_r, j_l}.$$

By Lemma A.8,

$$\mathbb{E}[T_z^2] \leq \frac{C_p}{t^r} \prod_{l=1}^r \|A^{z_l}\|_F^2 = \frac{C_p}{t^r} \prod_{l=1}^r \|A^{z_l}\|_{S_{2z_l}}^{2z_l},$$

where C_p is a constant that depends only on p . Overall, we have

$$\text{Var}(X) = \mathbb{E} \left[\left(\sum_z T_z \right)^2 \right] \leq C'_p \sum_z \mathbb{E}[T_z^2],$$

where the last inequality is by Cauchy-Schwartz and C'_p is the number of partitions of p . With the same argument as in Proposition 3.3, the lemma follows. \square

We thus arrive at the following theorem.

Theorem A.11. *For every $0 < \epsilon < 1/2$ and integer $p \geq 2$, there is a randomized one-pass streaming algorithm \mathcal{A} with space requirement $O(n^{2-4/p}/\epsilon^2)$, that given as input a PSD matrix $A \in \mathbb{R}^{n \times n}$, outputs with high probability a $(1 + \epsilon)$ -approximation of $\|A\|_{S_p}^p$. The algorithm processes an update in time $O(1/\epsilon^2)$, and computes the output (after the updates) in time $O(n^{(1-2/p)\omega})/\epsilon^2$, where $\omega < 3$ is the matrix multiplication constant.*

There is similarly a randomized $\lceil p/2 \rceil$ -pass streaming algorithm \mathcal{B} with space requirement $O(n^{1-1/(p-1)}/\epsilon^2)$, update time in a pass $O(1/\epsilon^2)$, and output time $O(n^{(1-2/p)}/\epsilon^2)$.

For even $p \geq 2$, both algorithms extend to general input $A \in \mathbb{R}^{n \times m}$ with $m \leq n$.

Proof. The algorithm and proof of correctness are almost identical to that of Theorem 3.2 and Theorem 3.5, with the exception that we replace the Gaussian sketching matrices with matrices drawn from $\mathcal{D}_{t,n}$. The correctness of this new sketch follows from Lemma A.10. Notice that every summand in the expansion of the variance expression involves at most 4 columns of any sketching matrix, and thus sketching matrices with 4-wise independent columns satisfy the variance upper bound we proved. This implies that $O_p(\log n)$ random bits suffice to generate the random matrices required in the algorithm.

Since the sketching matrices have only one non-zero entry per column, each update to the input matrix can be implemented in time $O(1)$ for each of the $O(1/\epsilon^2)$ independent copies of the basic sketch, thus the overall update time is $O(1/\epsilon^2)$. The computation of the output is straightforward. \square

B Unitarily Invariant Norms

B.1 Proofs of Theorem 4.1

The next lemma is an immediate consequence of [Woo14, Lemma 6.2].

Lemma B.1. *Let $S \in \mathbb{R}^{t \times d}$ be an (ϵ, δ, d) -OSE matrix. Let σ_i denote the i -th largest singular value of S . Then with probability at least $1 - \delta$,*

$$\forall i \in [d], \quad \sigma_i \in [1 - \epsilon, 1 + \epsilon].$$

Lemma 4.3. *Let $S \in \mathbb{R}^{t \times n}$ be an (ϵ, δ, d) -OSE and let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix of rank $r \leq d$. Then with probability at least $1 - \delta$, we have $\text{rank}(SAS^T) = r$ and there exists a permutation $\rho : [r] \rightarrow [r]$ such that*

$$\forall i \in [r], \quad \tilde{\lambda}_i(SAS^T) \in (1 \pm 3\epsilon)\tilde{\lambda}_{\rho(i)}(A),$$

where $\tilde{\lambda}_i(M)$ is the i th non-zero eigenvalue of M in decreasing order.

Proof. By the Spectral Theorem, we can write $A = P\Lambda P^T$, where $P \in \mathbb{R}^{n \times r}$ is a matrix with orthonormal columns and $\Lambda \in \mathbb{R}^{r \times r}$ is a diagonal matrix. The matrix $SP \in \mathbb{R}^{t \times r}$ is an (ϵ, δ, r) -OSE by Lemma 2.2. Lemma B.1 implies that, with probability at least $1 - \delta$, $\sigma_i(SP) \in (1 \pm \epsilon)$ for all $i \in [r]$. The lemma follows by applying Lemma 3.2 of [AN13]. \square

Lemma 4.4. Let $l(\cdot)$ be a unitarily invariant norm on matrices in $\mathbb{R}^{n \times n}$, and let $S \in \mathbb{R}^{t \times n}$ be an $(\epsilon, d\delta/n, d)$ -OSE matrix. Then for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, with probability at least $1 - \delta$,

$$l(SAS^T) \leq (1 + O(\epsilon)) \lceil n/d \rceil l(I_d) \|A\|_2,$$

where I_d is the identity matrix of order d .

Proof. We can write $A = U\Lambda U^T$ where U is an orthogonal matrix and Λ is a diagonal matrix. By Lemma 2.2, also SU is an $(\epsilon, d\delta/n, d')$ -OSE matrix, and thus it suffices to consider diagonal matrices, i.e., $A = \Lambda$ (the general case follows by replacing S throughout the proof with SU). For simplicity, we assume n is a multiple of d (the general case only requires to round n/d upwards), and write $\Lambda = \sum_{k=1}^{n/d} \Lambda^{(k)}$, where each matrix $\Lambda^{(k)}$ is obtained from Λ by breaking it into blocks of size $d \times d$ and zeroing all but the k -th main-diagonal block. Then by the triangle inequality,

$$l(S\Lambda S^T) = l\left(\sum_{k=1}^{n/d} S\Lambda^{(k)}S^T\right) \leq \sum_{k=1}^{n/d} l\left(S\Lambda^{(k)}S^T\right).$$

For each k , by Lemma 4.3, with probability at least $1 - d\delta/n$, the entire non-zero spectrum of $S\Lambda^{(k)}S^T$ approximates that of $\Lambda^{(k)}$ within factor $(1 \pm 3\epsilon)$. Now assuming this event occurs, we use (twice) the monotonicity of the norm $l(\cdot)$ in the singular values, and obtain

$$l(S\Lambda^{(k)}S^T) \leq (1 + 3\epsilon) l(\Lambda^{(k)}) \leq (1 + 3\epsilon) \|\Lambda\|_2 l(I_d).$$

The lemma follows by a union bound over these n/d events. \square

Proposition 4.5. If $l(\cdot)$ is a symmetric norm on \mathbb{R}^m , then

$$\begin{aligned} l(A) &= \max l(|u_1^T A v_1|, |u_2^T A v_2|, \dots, |u_m^T A v_m|)^T \\ \text{s.t. } u_1, \dots, u_m &\in \mathcal{S}^{m-1} \text{ are orthogonal} \\ v_1, \dots, v_m &\in \mathcal{S}^{n-1} \text{ are orthogonal} \end{aligned}$$

Proof. Let φ denote the maximum value attained in the expression on the right side above. Note that φ is well defined since the feasible region is compact. Let $A = UDV^T$ denote the s.v.d. of A with $U_1, \dots, U_m \in \mathcal{S}^{m-1}$ and $V_1, \dots, V_n \in \mathcal{S}^{n-1}$ the columns of U and V , respectively, and D a $m \times n$ diagonal matrix containing the singular values of A . Choosing $u_i = U_i$ and $v_i = V_i$, for all $i = 1, 2, \dots, m$, we see $\|A\| = \|\sigma\| \leq \varphi$, where σ denotes the m -dimensional vector of singular values of A .

For the reverse direction, let $(u_i)_{i \in [m]}, (v_i)_{i \in [m]}$ be some optimal solution. We decompose the vectors in the optimal solution as $u_i = \sum_{j=1}^m \alpha_{ij} U_j$ and $v_i = \sum_{j=1}^n \beta_{ij} V_j$, where $\sum_j \alpha_{ij}^2 = 1 = \sum_j \beta_{ij}^2$ for all i . Thus, $u_i^T A v_i = \sum_{j=1}^m \alpha_{ij} \beta_{ij} \sigma_j$. Furthermore, $\sum_j |\alpha_{ij} \beta_{ij}| \leq 1$ and $\sum_i |\alpha_{ij} \beta_{ij}| \leq 1$ by the Cauchy-Schwarz Inequality.

Let Λ be the $m \times m$ matrix with ij th entry $|\alpha_{ij} \beta_{ij}|$. The above inequalities imply that Λ is doubly substochastic, and by definition

$$(\Lambda \sigma)_i = \sum_j |\alpha_{ij} \beta_{ij}| \sigma_j \geq \sum_j \alpha_{ij} \beta_{ij} \sigma_j = u_i^T A v_i.$$

It is well known (Horn & Johnson, “Topics in Matrix Analysis” p.165) that Λ can be decomposed as a convex combination of partial permutation matrices, i.e. matrices form by zeroing out some entries of a permutation matrix. Thus, we may write a convex combination of permutation matrices $\Lambda' = \sum \lambda_k P_k \geq \Lambda$, where P_k are permutation matrices (augment the partial permutation matrices to permutation matrices arbitrarily) and the inequality holds coordinate-wise implying also that $\Lambda'\sigma \geq \Lambda\sigma$ coordinate-wise.

Now, by monotonicity of symmetric norms, the traingle inequality, and permutation symmetry we have

$$\varphi = \|(|u_1^T A v_1|, |u_2^T A v_2|, \dots, |u_m^T A v_m|)^T\| \leq \sum_k \lambda_k \|P_k \sigma\| = \|\sigma\| = \|A\|,$$

which is the desired inequality. \square

B.2 General Matrices and Q -norms

In this section we will show an even smaller sketch suffices when ℓ is a Q -norm.

Theorem B.2. *Let $A \in \mathbb{R}^{n \times n}$ be a PSD matrix. Let $\tilde{\ell} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}_{\geq 0}$ be a unitarily invariant norm and $\ell : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}_{\geq 0}$ be a Q -norm defined as $\ell(A) := \sqrt{\tilde{\ell}(AA^T)}$. Let $G \in \mathbb{R}^{t \times n}$ and $H \in \mathbb{R}^{t \times m}$ be two independent $(\epsilon, \delta, O(d \log n))$ -OSEs. With probability at least $1 - 2n\delta$,*

$$(1 - O(\epsilon))\ell(A)^2 \leq \tilde{\ell}(GAH^T H A^T G^T) \leq (1 + O(\epsilon))\ell(A)^2 + O\left(\frac{\ell(I_d)^2}{d} \sum_{i=d+1}^n \sigma_i(A)^2\right).$$

Proof of Theorem B.2. We will apply the inequalities in Theorem 4.1 twice. Suppose the inequalities hold with respect to the OSE G for $AH^T H A^T$ and with respect to the OSE H for AA^T . These both happen with probability at least $1 - 4n\delta$. First we have

$$\tilde{\ell}(GAH^T H A^T G^T) \geq (1 - O(\epsilon))\tilde{\ell}(AH^T H A^T) = (1 - O(\epsilon))\tilde{\ell}(H A A^T H^T) \geq (1 - O(\epsilon))\tilde{\ell}(AA^T).$$

Second,

$$\begin{aligned} (1 - O(\epsilon))\ell(A)^2 &\leq \tilde{\ell}(GAH^T H A^T G^T) \\ &\leq (1 + O(\epsilon)) \left(\tilde{\ell}(AH^T H A^T) + \frac{\tilde{\ell}(I_t)}{t} \sum_{i=d+1}^t \lambda_i(H A^T A H) \right). \end{aligned} \quad (16)$$

The first term on the right hand side of (16) can be bounded with Theorem 4.1 as

$$\tilde{\ell}(AH^T H A^T) \leq (1 + O(\epsilon))\tilde{\ell}(AA^T) + (1 + O(\epsilon)) \left(\frac{\tilde{\ell}(I_d)}{d} \sum_{i=d+1}^n \lambda_i(AA^T) \right).$$

It remains to bound the second term in (16), $\sum_{i=d+1}^t \lambda_i(H A^T A H)$. We will use a similar proof as used in Lemma 3.5 of [AN13], let Λ_l be the $n \times n$ diagonal matrix formed from eigenvalues of AA^T with diagonal entries $(\lambda_1, \dots, \lambda_d, 0, \dots, 0)$ and let Λ_s be the diagonal matrix with diagonal entries $(0, \dots, 0, \lambda_{d+1}, \dots, \lambda_n)$. Thus we can write $AA^T = U(\Lambda_l + \Lambda_s)U^T$. Recall that HU is still a (ϵ, δ, d) -OSE, thus, by Lidskii’s Inequality and Lemma A.1,

$$\sum_{i=d+1}^t \lambda_i(H A^T A H) \leq \text{Tr}(HU \Lambda_s U^T H^T) \leq (1 + \epsilon) \sum_{i=d+1}^n \sigma_i^2(A).$$

Thus, with probability at least $1 - O(\delta)$,

$$(1 - O(\epsilon))\ell(A)^2 \leq \tilde{\ell}(GAH^T HA^T G^T) \leq (1 + O(\epsilon))\ell(A)^2 + O\left(\frac{\tilde{\ell}(I_d)}{d} \sum_{i=d+1}^n \sigma_i(A)^2\right).$$

□

B.3 Lower Bounds

The complexity dependence of stable rank or intrinsic dimension (rather than the rank or the actual dimension) of a matrix allows us efficient algorithms on matrix norms of matrices that are actually full rank. In this section we will show that our algorithm is optimal (up to $\text{poly}(r)$ factor) in terms of the intrinsic dimension or stable rank of the matrix for all unitarily invariant norms that satisfy a condition, e.g. the Schatten- p norm with constant $p \notin \{1, 2\}$. matrix norms.

Theorem B.3. *Fix integer $0 \leq m \leq n$, Q -norm $\ell : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, if for all $t \in [n]$, $t/\ell(I_t)^2 \geq t^\alpha$ for some absolute constant $\alpha > 0$. Then any $O(1)$ -pass streaming algorithm that $(1 \pm \epsilon)$ -approximates ℓ on matrices of stable rank at most r requires $\Omega(\text{poly}(r))$ words of space.*

Proof. The proof follows from the vector norm lower bounds. In [BCKY15], the authors show that the $(O(1)$ -pass) streaming space complexity for vector Q -norm ℓ is $\tilde{\Theta}(t/\ell(I_t)^2)$. If there is an streaming algorithm for Q -norm for matrices, we can have an algorithm for the vector Q -norm by putting the vector on the diagonal of a all-zero matrix. For vector of dimension r , it is easy to see that the reduction matrix has dimension at most r , and thus of stable rank at most r . Therefore, the $(O(1)$ -pass) streaming space complexity of matrix Q -norm is $\tilde{\Theta}(r/\ell(I_r)^2) = \text{poly}(r)$. □

Theorem B.4. *Fix integer $0 \leq m \leq n$, constant $0 < \alpha < 1/2$, then there exists a unitarily invariant norm $\ell : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, such that for all $t \in [n]$, $t/\ell(I_t) \geq t^\alpha$ and any linear-sketch based streaming algorithm that $(1 \pm \epsilon)$ -approximates ℓ on PSD matrices of intrinsic dimension at most r requires $\Omega(\text{poly}(r))$ words of space.*

This theorem follows from the following lemma since a Schatten- p norm for $1 < p < 2$ satisfies the conditions of the above lemma.

Lemma B.5 (A consequence of [LNW14]). *Suppose $X \in \mathbb{R}^{n \times n}$ is a PSD matrix and $1 < p < 2$. Suppose that an algorithm takes k linear sketches of X and computes Y with $(1 - c_p)\|X\|_{S_p}^p \leq Y \leq (1 + c_p)\|X\|_{S_p}^p$ with probability at least $3/4$ then $k = \Omega(\sqrt{n})$, where c_p is a constant depends only on p .*

Proof. The proof of Theorem 5.4 of [LNW14] does not explicitly state the result for PSD matrices. But we can symmetrize the hard instance matrices as follows. Suppose the non-symmetric hard instance matrix as $B \in \mathbb{R}^n$, then define matrix

$$A = aI_{2n} + \begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix},$$

where $a > 1$ is a constant such that with probability at least 0.99, A is a PSD matrix (when B is drawn from the hard distribution). The hard distribution is: case 1, $B = (G, GM)$, where $G \in \mathbb{R}^{n \times n/2}$ is a column normalized Gaussian matrix, and $M \sim O_{n/2}$; case 2, $B = G'$, where

$G' \in \mathbb{R}^{n \times n}$ is a column normalized Gaussian matrix. For both cases, with high probability every non-zero singular values is $\Theta(1)$. Thus it is suffice to set $\alpha = \Theta(1)$. Now the norm of A of both distributions can be computed exactly using the same formula as (H.15) and (3.4), and replace their I_p and J_p as follows i.e.

$$I'_p = \int_0^4 (a+x)^{p/2} \cdot \frac{\sqrt{(4-x)x}}{2\pi x} + \int_0^4 (a-x)^{p/2} \cdot \frac{\sqrt{(4-x)x}}{2\pi x},$$

and

$$J'_p = 2^{p/2} \int_0^4 (a+x)^{p/2} \cdot \frac{(b-x)(x-a)}{\pi x} + 2^{p/2} \int_0^4 (a-x)^{p/2} \cdot \frac{(b-x)(x-a)}{\pi x}.$$

Other part of the proof shall follow directly. \square

C Proofs of Section 5

Lemma C.1. *For $n \geq 1$, let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $\lambda_i^*(A)$ be the i -th largest eigenvalue of A in absolute value. Fix $\epsilon, \phi \in (0, 1)$. Let $G \in \mathbb{R}^{t \times n}$ be both $(O(\epsilon), 1/\text{poly } n, \phi^{-1})$ -OSE and $(0.1, 1/\text{poly } n, \phi^{-1}\epsilon^{-2})$ -OSE. Then with probability at least $9/10$, the (ϕ^{-1}) -largest eigenvalues of GAG^T in absolute value are $1 \pm \epsilon$ multiplicative and $O(\epsilon^2 \phi \sum_{i=1/\phi+1}^n |\lambda_i^*(A)| + |\lambda_{1/\phi}^*(A)|)$ additive error approximations of the ϕ^{-1} largest eigenvalues of A in absolute value (under some mapping between the two sets of eigenvalues).*

Since the proof is similar, we only point out the necessary changes to the proof of [AN13, Lemma 3.4]. For the full proof, please refer to proof of [AN13, Lemma 3.4].

Proof. By the Spectral Theorem, we can write $A = U^T \Lambda U$, where U is an orthormal matrix, and $\Lambda = \text{diag}(\lambda_1^*, \dots, \lambda_n^*)$ such that $|\lambda_1^*| \geq |\lambda_2^*| \geq \dots |\lambda_n^*|$. By Lemma 2.2, also GU is an $(O(\epsilon), 1/\text{poly } n, \phi^{-1})$ -OSE matrix (with the same parameters), and thus it suffices to consider a diagonal matrix $A = \Lambda$. Now let Λ_i ($i \geq 1$) be the diagonal matrices with eigenvalues of absolute value in the range $[|\lambda_{1/\phi}^*(A)|2^{-i}, |\lambda_{1/\phi}^*(A)|2^{-i+1})$ and Λ_0 be the diagonal matrix of the first $1/\phi$ eigenvalues. Denote $\Lambda = \sum_{i=0}^{\infty} \Lambda_i$. By Lemma 4.3, the non-zero eigenvalues of $G^T \Lambda_0 G$ are $(1 \pm \epsilon)$ approximation to those of Λ_0 . For the rest of the proof to work, we only need to show the additive error term $\sum_{i \geq 1} \|G \Lambda_i G^T\|_2$ is small. And it is suffice to show that, with probability at least $1 - 1/\text{poly}(n)$, for all diagonal matrix D with diagonal entries from $\{0, 1\}$, $\|GDG^T\|_2 \leq O(\max(\epsilon^2 \phi \text{Tr}(D), 1))$. We use the fact that GU is a $(0.1, 1/\text{poly } n, \phi^{-1}\epsilon^{-2})$ -OSE. By Lemma 4.4 we have that, with probability at least $1 - 1/\text{poly } n$,

$$\|GDG^T\|_2 \leq O(\lceil \epsilon^2 \phi \text{Tr}(D) \rceil) = O(\max(\epsilon^2 \phi \text{Tr}(D), 1)).$$

\square

Lemma C.2. *For $n, m \geq 1$, let $A \in \mathbb{R}^{n \times m}$ be an real matrix and $s_i(A)$ be the i -th largest singular value of A . Fix $\epsilon, \phi \in (0, 1)$. Let $G \in \mathbb{R}^{t \times n}, H \in \mathbb{R}^{t \times m}$ be independent OSE matrices that are both of $(O(\epsilon), 1/\text{poly } n, (\phi^{-1} + 1))$ -OSE and of $(0.1, 1/\text{poly } n, (\phi^{-1} + 1)\epsilon^{-2})$ -OSE. Then with probability at least $9/10$, the ϕ^{-1} -largest singular values of GAH^T are $1 \pm O(\epsilon)$ multiplicative and $O(\epsilon^2 \phi \sum_{i=1/\phi+1}^n s_i^2(A) + s_{1/\phi}^2(A))$ additive error approximations of the ϕ^{-1} largest singular values of A .*

We point out the necessary changes to the proof of [AN13, Lemma 3.5]. For the full proof, please refer to [AN13, Lemma 3.5].

Proof. First note that $A = U\Lambda V^T$, where $\Lambda = \text{diag}(s_1, s_2, \dots, s_n)$ be the diagonal matrix of singular values. Note that $s_1 \geq s_2 \geq \dots s_n \geq 0$. Now we write $\Lambda = \Lambda_l + \Lambda_s$, where Λ_l contains the top ϕ^{-1} singular values and Λ_s contains the rest. By Lemma C.1, we have that ϕ^{-1} eigenvalues of $GAH^T HA^T G^T$ are $(1 \pm \epsilon)$ approximations to those of $AH^T HA^T$ and up to additive error $O(\epsilon^2 \phi \sum_{i=\phi^{-1}+1}^n \lambda_i(AH^T HA^T) + \lambda_{\phi^{-1}}(AH^T HA^T))$. Since $AH^T HA^T$ has the same set of eigenvalues of $HAA^T H^T$. Thus the top ϕ^{-1} eigenvalues of $AH^T HA^T$ are $(1 \pm \epsilon)$ multiplicative approximation to those of $A^T A$ up to $O(\epsilon^2 \phi \sum_{i=\phi^{-1}+1}^n \lambda_i(A^T A) + \lambda_{\phi^{-1}}(A^T A))$ additive error. Therefore, it remains to show that $O(\epsilon^2 \phi \sum_{i=\phi^{-1}+1}^n \lambda_i(AH^T HA^T) + \lambda_{\phi^{-1}}(AH^T HA^T))$ is upper bounded by $O(\epsilon^2 \phi \sum_{i=\phi^{-1}+1}^n \lambda_i(A^T A) + \lambda_{\phi^{-1}}(A^T A))$. First note that, by Lemma C.1, with probability at least $1 - 1/\text{poly } n$, $\lambda_{\phi^{-1}}(AH^T HA^T) = O(\epsilon^2 \phi \sum_{i=\phi^{-1}+1}^n \lambda_i(A^T A) + \lambda_{\phi^{-1}}(A^T A))$. By Lidskii inequality and Lemma A.1, with probability at least $1 - 1/\text{poly } n$,

$$\begin{aligned}
\sum_{i=\phi^{-1}+1}^n \lambda_i(AH^T HA^T) &= \sum_{i=\phi^{-1}+1}^n \lambda_i(HA^T AH^T) \\
&= \sum_{i=\phi^{-1}+1}^n \lambda_i(HV(\Lambda_l^2 + \Lambda_s^2)V^T H^T) \\
&\leq \text{Tr}(HV\Lambda_s^2 V^T H^T) \\
&\leq (1 + O(\epsilon)) \text{Tr}(\Lambda_s^2).
\end{aligned} \tag{17}$$

□